Published in final edited form as:
J Mach Learn Res. 2012 June 1; 13: 1839–1864.

Estimation and Selection via Absolute Penalized Convex Minimization And Its Multistage Adaptive Applications

Jian Huang and
Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242, USA

Cun-Hui Zhang
Department of Statistics and Biostatistics, Rutgers University, Piscataway, New Jersey 08854, USA

Jian Huang: JIAN-HUANG@UIOWA.EDU; Cun-Hui Zhang: CZHANG@STAT.RUTGERS.EDU

Abstract

The ℓ_1-penalized method, or the Lasso, has emerged as an important tool for the analysis of large data sets. Many important results have been obtained for the Lasso in linear regression which have led to a deeper understanding of high-dimensional statistical problems. In this article, we consider a class of weighted ℓ_1-penalized estimators for convex loss functions of a general form, including the generalized linear models. We study the estimation, prediction, selection and sparsity properties of the weighted ℓ_1-penalized estimator in sparse, high-dimensional settings where the number of predictors p can be much larger than the sample size n. Adaptive Lasso is considered as a special case. A multistage method is developed to approximate concave regularized estimation by applying an adaptive Lasso recursively. We provide prediction and estimation oracle inequalities for single- and multi-stage estimators, a general selection consistency theorem, and an upper bound for the dimension of the Lasso estimator. Important models including the linear regression, logistic regression and log-linear models are used throughout to illustrate the applications of the general results.

Keywords

variable selection; penalized estimation; oracle inequality; generalized linear models; selection consistency; sparsity

1. Introduction

High-dimensional data arise in many diverse fields of scientific research. For example, in genetic and genomic studies, more and more large data sets are being generated with rapid advances in biotechnology, where the total number of variables p is larger than the sample size n. Fortunately, statistical analysis is still possible for a substantial subset of such problems with a sparse underlying model where the number of important variables is much smaller than the sample size. A fundamental problem in the analysis of such data is to find reasonably accurate sparse solutions that are easy to interpret and can be used for the prediction and estimation of covariable effects. The ℓ_1-penalized method, or the Lasso (Tibshirani, 1996; Chen et al., 1998), has emerged as an important approach to finding such solutions in sparse, high-dimensional statistical problems.

In the last few years, considerable progress has been made in understanding the theoretical properties of the Lasso in $p \gg n$ settings. Most results have been obtained for linear regression models with a quadratic loss. Greenshtein and Ritov (2004) studied the prediction performance of the Lasso in high-dimensional least squares regression. Meinshausen and

Bühlmann (2006) showed that, for neighborhood selection in the Gaussian graphical models, under a neighborhood stability condition on the design matrix and certain additional regularity conditions, the Lasso is selection consistent even when $p \to \infty$ at a rate faster than n. Zhao and Yu (2006) formalized the neighborhood stability condition in the context of linear regression as a strong irrepresentable condition. Candes and Tao (2007) derived an upper bound for the ℓ_2 loss of a closely related Dantzig selector in the estimation of regression coefficients under a condition on the number of nonzero coefficients and a uniform uncertainty principle on the design matrix. Similar results have been obtained for the Lasso. For example, upper bounds for the ℓ_q loss of the Lasso estimator has being established by Bunea et al. (2007) for $q = 1$, Zhang and Huang (2008) for $q \in [1;2]$, Meinshausen and Yu (2009) for $q = 2$, Bickel et al. (2009) for $q \in [1;2]$, and Zhang (2009) and Ye and Zhang (2010) for general $q \geq 1$. For convex minimization methods beyond linear regression, van de Geer (2008) studied the Lasso in high-dimensional generalized linear models (GLM) and obtained prediction and ℓ_1 estimation error bounds. Negahban et al. (2010) studied penalized M-estimators with a general class of regularizers, including an ℓ_2 error bound for the Lasso in GLM under a restricted strong convexity and other regularity conditions.

Theoretical studies of the Lasso have revealed that it may not perform well for the purpose of variable selection, since its required irrepresentable condition is not properly scaled in the number of relevant variables. In a number of simulation studies, the Lasso has shown weakness in variable selection when the number of nonzero regression coefficients increases. As a remedy, a number of proposals have been introduced in the literature and proven to be variable selection consistent under regularity conditions of milder forms, including concave penalized LSE (Fan and Li, 2001; Zhang, 2010a), adaptive Lasso (Zou, 2006; Meier and Bühlmann, 2007; Huang et al., 2008), stepwise regression (Zhang, 2011a), and multi-stage methods (Hunter and Li, 2005; Zou and Li, 2008; Zhang, 2010b, 2011b).

In this article, we study a class of weighted ℓ_1-penalized estimators with a convex loss function. This class includes the Lasso, adaptive Lasso and multistage recursive application of adaptive Lasso in generalized linear models as special cases. We study prediction, estimation, selection and sparsity properties of the weighted ℓ_1-penalized estimator based on a convex loss in sparse, high-dimensional settings where the number of predictors p can be much larger than the sample size n. The main contributions of this work are as follows.

- We extend the existing theory for the unweighted Lasso from linear regression to more general convex loss function.
- We develop a multistage method to approximate concave regularized convex minimization with recursive application of adaptive Lasso, and provide sharper risk bounds for this concave regularization approach in the general setting.
- We apply our results to a number of important special cases, including the linear, logistic and log-linear regression models.

This article is organized as follows. In Section 2 we describe a general formulation of the absolute penalized minimization problem with a convex loss, along with two basic inequalities and a number of examples. In Section 3 we develop oracle inequalities for the weighted Lasso estimator for general quasi star-shaped loss functions and an ℓ_2 bound on the prediction error. In Section 4 we develop multistage recursive applications of adaptive Lasso as an approximate concave regularization method and provide sharper oracle inequalities for this approach. In Section 5 we derive sufficient conditions for selection consistency. In Section 6 we provide an upper bound on the dimension of the Lasso estimator. Concluding remarks are given in Section 7. All proofs are provided in an appendix.

2. Absolute Penalized Convex Minimization

In this section, we define the weighted Lasso for a convex loss function and characterize its solutions via the KKT conditions. We then derive some basic inequalities for the weighted Lasso solutions in terms of the symmetrized Bregman divergence (Bregman, 1967; Nielsen and Nock, 2007). We also illustrate the applications of the basic inequalities in several important examples.

2.1 Definition and the KKT Conditions

We consider a general convex loss function of the form

$$\ell(\beta)=\psi(\beta)-\langle\beta,z\rangle, \quad (1)$$

where $\psi(\beta)$ is a known convex function, z is observed, and β is unknown. Unless otherwise stated, the inner product space is $\mathbb{R}^p$, so that $\{z,\beta\}\subset\mathbb{R}^p$ and $\langle\beta,z\rangle=\beta'z$. Our analysis of (1) requires certain smoothness of the function $\psi(\beta)$ in terms of its differentiability. In what follows, such smoothness assumptions are always explicitly described by invoking the derivative of ψ. For any $v=(v_1,\ldots,v_p)'$, we use $\|v\|$ to denote a general norm of v and $|v|_q$ the ℓ_q norm $(\Sigma_j|v_j|^q)^{1/q}$, with $|v|_\infty=\max_j|v_j|$. Let $\hat{w}\in\mathbb{R}^p$ be a (possibly estimated) weight vector with nonnegative elements $\hat{w}_j$, $1\le j\le p$, and $\hat{W}=\mathrm{diag}(\hat{w})$. The weighted absolute penalized estimator, or weighted Lasso, is defined as

$$\hat{\beta}=\arg\min_{\beta}\left\{\ell(\beta)+\lambda|\hat{W}\beta|_1\right\}. \quad (2)$$

Here we focus on the case where $\hat{W}$ is diagonal. In linear regression, Tibshirani and Taylor (2011) considered non-diagonal, predetermined $\hat{W}$ and derived an algorithm for computing the solution paths.

A vector $\hat{\beta}$ is a global minimizer in (2) if and only if the negative gradient at $\hat{\beta}$ satisfies the Karush-Kuhn-Tucker (KKT) conditions,

$$g=-\dot{\ell}(\hat{\beta})=z-\dot{\psi}(\hat{\beta}),\begin{cases} g_j=\hat{w}_j\lambda\,\mathrm{sgn}(\hat{\beta}_j) & \text{if }\hat{\beta}_j\neq 0\\ g_j\in\hat{w}_j\lambda[-1,1] & \text{all } j,\end{cases} \quad (3)$$

where $\dot{\ell}(\beta)=(\partial/\partial\beta)\ell(\beta)$ and $\dot{\psi}(\beta)=(\partial/\partial\beta)\psi(\beta)$. Since the KKT conditions are necessary and sufficient for (2), results on the performance of $\hat{\beta}$ can be viewed as analytical consequences of (3).

The estimator (2) includes the ℓ_1-penalized estimator, or the Lasso, with the choice $\hat{w}_j=1$, $1\le j\le p$. A careful study of the (unweighted) Lasso in general convex minimization (1) is by itself an interesting and important problem. Our work includes the Lasso as a special case since $\hat{w}_j=1$ is allowed in our theorems.

In practice, unequal $\hat{w}_j$ arise in many ways. In adaptive Lasso (Zou, 2006), a decreasing function of a certain initial estimator of β_j is used as the weight $\hat{w}_j$ to remove the bias of the Lasso. In Zou and Li (2008) and Zhang (2010b), the weights $\hat{w}_j$ are computed iteratively with $\hat{w}_j=\dot{\rho}_\lambda(\hat{\beta}_j)$, where $\dot{\rho}_\lambda(t)=(d/dt)\rho_\lambda(t)$ with a suitable concave penalty function $\rho_\lambda(t)$. This is also designed to remove the bias of the Lasso, since the concavity of $\rho_\lambda(t)$ guarantees smaller weight for larger $\hat{\beta}_j$. In Section 4, we provide results on the improvements of this weighted Lasso over the standard Lasso. In linear regression, Zhang (2010b) gave sufficient

conditions under which this iterative algorithm provides smaller weights $\hat{w}_j$ for most large β_j. Such nearly unbiased methods are expected to produce better results than the Lasso when a significant fraction of nonzero $|\beta_j|$ are of the order λ or larger. Regardless of the computational methods, the results in this paper demonstrate the benefits of using data dependent weights in a general class of problems with convex losses.

Unequal weights may also arise for computational reasons. The Lasso with $\hat{w}_j = 1$ is expected to perform similarly to weighted Lasso with data dependent $1 \le \hat{w}_j \le C_0$, with a fixed C_0. However, the weighted Lasso is easier to compute since $\hat{w}_j$ can be determined as a part of an iterative algorithm. For example, in a gradient descent algorithm, one may take larger steps and stop the computation as soon as the KKT conditions (3) are attained for any weights satisfying $1 \le \hat{w}_j \le C_0$.

The weight function $\hat{w}_j$ can be also used to standardize the penalty level, for example with $\hat{w}_j = \{\ddot{\psi}_{j\,j}(\hat{\beta})\}^{1/2}$, where $\ddot{\psi}_{j\,j}(\beta)$ is the j-th diagonal element of the Hessian matrix of $\psi(\beta)$. When $\psi(\beta)$ is quadratic, for example in linear regression, $\hat{w}_j = \{\ddot{\psi}_{jj}(\hat{\beta})\}^{1/2}$ does not depend on $\hat{\beta}$. However, in other convex minimization problems, such weights need to be computed iteratively.

Finally, in certain applications, the effects of a certain set S_* of variables are of primary interest, so that penalization of $\hat{\beta}_{S*}$, and thus the resulting bias, should be avoided. This leads to "semi-penalized" estimators with $\hat{w}_j = 0$ for $j \in S_*$, for example, with weights $\hat{w}_i = I\{j \notin S_*\}$.

2.2 Basic Inequalities, Prediction, and Bregman Divergence

Let β^* denote a target vector for β. In high-dimensional models, the performance of an estimator $\hat{\beta}$ is typically measured by its proximity to a target under conditions on the sparsity of β^* and the size of the negative gradient $-\dot{\ell}(\beta^*) = z - \dot{\psi}(\beta^*)$. For ℓ_1-penalized estimators, such results are often derived from the KKT conditions (3) via certain basic inequalities, which are direct consequences of the KKT conditions and have appeared in different forms in the literature, for example, in the papers cited in Section 1. Let $D(\beta,\beta^*) = \ell(\beta) - \ell(\beta^*) - \langle \dot{\ell}(\beta^*), \beta - \beta^* \rangle$ be the Bregman divergence (Bregman, 1967) and consider its symmetrized version (Nielsen and Nock, 2007)

$$\Delta(\beta,\beta^*) = D(\beta,\beta^*) + D(\beta^*,\beta) = \langle \beta - \beta^*, \dot{\psi}(\beta) - \dot{\psi}(\beta^*) \rangle. \quad (4)$$

Since ψ is convex, $\Delta(\beta, \beta^*) \ge 0$. Two basic inequalities below provide upper bounds for the symmetrized Bregman divergence $\Delta(\hat{\beta}, \beta^*)$. The sparsity of β^* is measured by a weighted ℓ_1 norm of β^* in the first one and by a sparse set in the second one.

Let S be any set of indices satisfying $S \supseteq \{j : \beta_j^* \neq 0\}$ and let S^c be the complement of S in $\{1, \ldots, p\}$. We shall refer to S as the sparse set. Let $W = \mathrm{diag}(w)$ for a possibly unknown vector $w \in \mathbb{R}^p$ with elements $w_j \ge 0$. Define

$$z_0^* = |\{z - \dot{\psi}(\beta^*)\}_S|_\infty, \quad z_1^* = |W_{S^c}^{-1}\{z - \dot{\psi}(\beta^*)\}_{S^c}|_\infty, \quad (5)$$

$$\Omega_0 = \{\hat{w}_j \le w_j \forall j \in S\} \cap \{w_j \le \hat{w}_j \forall j \in S^c\}, \quad (6)$$

where for any p-vector v and set A, $v_A = (v_j : j \in A)'$. Here and in the sequel M_{AB} denotes the $A \times B$ subblock of a matrix M and $M_A = M_{AA}$.

Lemma 1

i. *Let* β^* *be a target vector. In the event* $\Omega_0 \cap \{|(z - \dot{\psi}(\beta^*))_j| \le \hat{w}_j\lambda \ \forall j\}$,

$$\Delta(\widehat{\beta}, \beta^*) \le 2\lambda|\widehat{W}\beta^*|_1 \le 2\lambda|W\beta^*|_1. \quad (7)$$

ii. *For any target vector* β^* *and* $S \supseteq \{j:\beta_j^* \ne 0\}$, *the error* $h = \hat{\beta} - \beta^*$ *satisfies*

$$\begin{aligned}\Delta(\beta^*+h, \beta^*)+(\lambda-z_1^*)|W_{S^c}h_{S^c}|_1 &\le \langle h_S, g_S-\{z-\dot{\psi}(\beta^*)\}_S\rangle \\ &\le (|w_S|_\infty\lambda+z_0^*)|h_S|_1\end{aligned} \quad (8)$$

in Ω_0 *for a certain negative gradient vector g satisfying* $|g_j| \le \hat{w}_j\lambda$. *Consequently, in* $\Omega_0 \cap \{(|w_S|_\infty\lambda+z_0^*)/(\lambda-z_1^*) \le \xi\}$, $h \ne 0$ *belongs to the sign-restricted cone* $\mathscr{C}(\xi, S) = \{b \in \mathscr{C}(\xi, S) : b_j(\psi(\beta+ b) - \psi(\beta))_j \le 0 \ \forall\, j \in S^c\}$, *where*

$$\mathscr{C}(\xi, S)=\{b \in \mathbb{R}^p:|W_{S^c}b_{S^c}|_1 \le \xi|b_S|_1 \ne 0\}. \quad (9)$$

Remark 2: *Sufficient conditions are given in Subsection 3.2 for* $\{|(z - \dot{\psi}(\beta^*))_j| \le \hat{w}_j\lambda \ \forall j\}$ *to hold with high probability in generalized linear models. See Lemma 8, Remarks 10 and 11 and Examples 7, 8, and 9.*

A useful feature of Lemma 1 is the explicit statements of the monotonicity of the basic inequality in the weights. By Lemma 1 (ii), it suffices to study the analytical properties of the penalized criterion with the error $h = \hat{\beta} - \beta^*$ in the sign-restricted cone, provided that the event $(|w_S|_\infty\lambda+z_0^*)/(\lambda-z_1^*) \le \xi$ has large probability. However, unless $\mathscr{C}(\xi, S)$ is specified, we will consider the larger cone in (9) in order to simplify the analysis. The choices of the target vector β^*, the sparse set $S \supseteq \{j:\beta_j^* \ne 0\}$, weight vector $\hat{w}$ and its bound w are quite flexible. The main requirement is that $\{|S|, z_0^*, z_1^*\}$ should be small. In linear regression or generalized linear models, we may conveniently consider β^* as the vector of true regression coefficients under a probability measure P_{β^*}. However, β^* can also be a sparse version of a true β, for example, $\beta_j^*=\beta_j I\{|\beta_j| \ge \tau\}$ for a threshold value τ under P_β.

The upper bound in Lemma 1 (i) gives the so called "slow rate" of convergence for the Bregman divergence. In Section 3, we provide "fast rate" of convergence for the Bregman divergence via oracle inequalities for $|h_S|_1$ in (8).

The symmetrized Bregman divergence $\Delta(\hat{\beta}, \beta^*)$ has the interpretations as the regret in prediction error in linear regression, the symmetrized Kullback-Leibler (KL) divergence in generalized linear models (GLM) and density estimation, and a spectrum loss for the graphical Lasso, as shown in examples below. These quantities can be all viewed as the size of the prediction error since they measure distances between a target density of the observations and an estimated density.

Example 1 (Linear regression): Consider the linear regression model

$$y_i=\sum_{j=1}^{p}x_{ij}\beta_j+\varepsilon_i, \ \ i=1,\ldots,n, \quad (10)$$

where y_i is the response variable, x_{ij} are predictors or design variables, and ε_i is the error term. Let $y = (y_1, \ldots, y_n)'$ and let X be the design matrix whose ith row is $x^i = (x_{i1}, \ldots, x_{ip})$. The estimator (2) *can be written as a weighted Lasso with* $\psi(\beta) = |X\beta|_2^2/(2n)$ *and* $z = X'y/n$ *in* (1). *For predicting a vector* $\tilde{y}$ *with* $\mathrm{E}_{\beta^*}[\tilde{y}|X, y] = X\beta^*$,

$$n\Delta(\widehat{\beta}, \beta^*) = |X\widehat{\beta} - X\beta^*|_2^2 = \mathrm{E}_{\beta^*}[\,|\tilde{y} - X\widehat{\beta}|_2^2 | X, y] - \min_{\delta(X,y)} \mathrm{E}_{\beta^*}[\,|\tilde{y} - \delta(X,y)|_2^2 | X, y]$$

is the regret of using the linear predictor $X\hat{\beta}$ *compared with the optimal predictor. See* Greenshtein and Ritov (2004) *for several implications of* (7).

Example 2 (Logistic regression): *We observe* $(X, y) \in \mathbb{R}^{n\times(p+1)}$ *with independent rows* (x^i, y_i), *where* $y_i \in \{0,1\}$ *are binary response variables with*

$$\mathrm{P}_\beta(y_i = 1|x^i) = \pi_i(\beta) = \exp(x^i\beta)/(1+\exp(x^i\beta)), 1 \le i \le n. \quad (11)$$

The loss function (1) is the average negative log-likelihood:

$$\ell(\beta) = \psi(\beta) - z'\beta \text{ with } \psi(\beta) = \sum_{i=1}^{n} \frac{\log(1+\exp(x^i\beta))}{n}, z = X'y/n. \quad (12)$$

Thus, (2) *is a weighted* ℓ_1 *penalized MLE. For probabilities* $\{\pi', \pi''\} \subset (0,1)$, *the KL information is* $K(\pi', \pi'') = \pi'\log(\pi'/\pi'') + (1-\pi')\log\{(1-\pi')/(1-\pi'')\}$. *Since* $\dot{\psi}(\beta) = \sum_{i=1}^{n} x^i\pi_i(\beta)/n$ *and logit*$(\pi_i(\beta^*)) - $ *logit*$(\pi_i(\beta)) = x^i(\beta^* - \beta)$, (4) *gives*

$$\Delta(\beta, \beta^*) = \frac{1}{n}\sum_{i=1}^{n} \{K(\pi_i(\beta^*), \pi_i(\beta)) + K(\pi_i(\beta), \pi_i(\beta^*))\}.$$

Thus, the symmetrized Bregman divergence $\Delta(\beta^*, \beta)$ *is the symmetrised KL-divergence.*

Example 3 (GLM): *The GLM contains the linear and logistic regression models as special cases. We observe* $(X, y) \in \mathbb{R}^{n\times(p+1)}$ *with rows* (x^i, y_i). *Suppose that conditionally on* X, y_i *are independent under* P_β *with*

$$y_i \sim f(y_i|\theta_i) = \exp\left(\frac{\theta_i y_i - \psi_0(\theta_i)}{\sigma^2} + \frac{c(y_i, \sigma)}{\sigma^2}\right), \theta_i = x^i\beta. \quad (13)$$

Let $f_{(n)}(y|X, \beta) = \prod_{i=1}^{n} f(y_i|x^i\beta)$. *The loss function can be written as a normalized negative likelihood* $\ell(\beta) = (\sigma^2/n)\log f_{(n)}(y|X, \beta)$ *with* $\psi(\beta) = \sum_{i=1}^{n}\{\psi_0(x^i\beta) + c(y_i, \sigma)\}/n$ *and* $z = X'y/n$. *The KL divergence is*

$$D\ (f_n(\cdot|X,\beta^*)\|f_n(\cdot|X,\beta)) = \mathrm{E}_{\beta^*}\log\left(\frac{f_{(n)}(y|X,\beta^*)}{f_{(n)}(y|X,\beta)}\right).$$

The symmetrized Bregman divergence can be written as

$$\Delta(\widehat{\beta},\beta^*) = \frac{\sigma^2}{n}\left\{D\left(f_{(n)}(\cdot|X,\beta^*)\|f_{(n)}(\cdot|X,\widehat{\beta})\right) + D\left(f_{(n)}(\cdot|X,\widehat{\beta})\|f_{(n)}(\cdot|X,\beta^*)\right)\right\}.$$

Example 4 (Nonparametric density estimation): *Although the focus of this paper is on regression models, here we illustrate that* $\Delta(\beta,\hat{\beta}^*)$ *is the symmetrised KL divergence in the context of non-parametric density estimation. Suppose the observations* $y = (y_1, \ldots, y_n)'$ *are iid from* $f(\cdot|\beta) = \exp\{\langle\beta, T(\cdot)\rangle - \psi(\beta)\}$ *under* P_β*, where* $T(\cdot) = (u_j(\cdot), j \le p)'$ *with certain basis functions* $u_j(\cdot)$*. Let the loss function* $\ell(\beta)$ *in (1) be the average negative log-likelihood* $n^{-1}\sum_{i=1}^n \log f(y_i|\beta)$ *with* $z = n^{-1}\sum_{i=1}^n T(y_i)$*.* Since $\mathrm{E}_\beta T(y_i) = \dot{\psi}(\beta)$*, the KL divergence is*

$$D\ (f(\cdot|\beta^*)\|f(\cdot|\beta)) = \mathrm{E}_{\beta^*}\log\left(\frac{f(y_i|\beta^*)}{f(y_i|\beta)}\right) = \psi(\beta) - \psi(\beta^*) - \langle\beta-\beta^*, \psi(\beta^*)\rangle.$$

Again, the symmetrized Bregman divergence is the symmetrised KL divergence between the target density $f(\cdot|\beta^*)$ *and the estimated density* $f(\cdot|\hat{\beta})$*:*

$$\Delta(\beta,\beta^*) = D\left(f(\cdot|\beta^*)\|f(\cdot|\widehat{\beta})\right) + D\left(f(\cdot|\widehat{\beta})\|f(\cdot|\beta^*)\right).$$

van de Geer (2008) *pointed out that for this example, the natural choices of the basis functions* u_j *and weights* w_j *satisfy* $\int u_j d\nu = 0$ *and* $w_k^2 = \int u_k^2 d\nu$*.*

Example 5 (Graphical Lasso): *Suppose we observe* $X \in \mathbb{R}^{n\times p}$ *and would like to estimate the precision matrix* $\beta = (EX'X/n)^{-1} \in \mathbb{R}^{p\times p}$*. In the graphical Lasso, (1) is the length normalized negative likelihood with* $\psi(\beta) = -\log\det\beta$*,* $z = -X'X/n$*, and* $\langle\beta, z\rangle = -\mathrm{trace}(\beta z)$*. Since the gradient of* ψ *is* $\dot{\psi}(\beta) = \mathrm{E}_\beta z = -\beta^{-1}$*, we find*

$$\Delta(\beta,\beta^*) = \mathrm{trace}\left((\widehat{\beta}-\beta^*)\left((\beta^*)^{-1}-\widehat{\beta}^{-1}\right)\right) = \sum_{j=1}^p (\lambda_j - 1)^2/\lambda_j,$$

where $(\lambda_1, \ldots, \lambda_p)$ *are the eigenvalues of* $(\beta^*)^{-1/2}\hat{\beta}(\beta^*)^{-1/2}$*. In graphical Lasso, the diagonal elements are typically not penalized. Consider* $\hat{w}_{jk} = I\{j \ne k\}$*, so that the penalty for the off-diagonal elements are uniformly weighted. Since Lemma 1 requires* $|(z - \dot{\psi}(\beta^*))_{jk}| \le \hat{w}_{jk}\lambda$*,* β^* *is taken to match* $X'X/n$ *on the diagonal and the true* β *in correlations. Let* $S = \{(j,k) : \beta_{jk} \ne 0, j \ne k\}$*. In the event* $\max_{j\ne k}|z_{jk} - \beta^*_{jk}| \le \lambda$*, Lemma 1 (i) gives*

$$|S|\lambda\max_{j\neq k}|\beta^*_{jk}|=o(1) \Rightarrow \|(\beta^*)^{-1/2}\widehat{\beta}(\beta^*)^{-1/2}-I_{p\times p}\|_2=o(1)$$

where $\|\cdot\|_2$ *is the spectrum norm.* Rothman et al. (2008) *proved the consistency of the graphical Lasso under similar conditions with a different analysis.*

3. Oracle Inequalities

In this section, we extract upper bounds for the estimation error $\hat{\beta} - \beta^*$ from the basic inequality (8). Since (8) is monotone in the weights, the oracle inequalities are sharper when the weights $\hat{w}_j$ are smaller in $S \supseteq \{j:\beta^*_j \neq 0\}$ and larger in S^c.

We say that a function $\phi(b)$ defined in $\mathbb{R}^p$ is quasi star-shaped if $\phi(tb)$ is continuous and non-decreasing in $t \in [0,\infty)$ for all $b \in \mathbb{R}^p$ and $\lim_{b\to 0}\phi(b) = 0$. All seminorms are quasi star-shaped. The sublevel sets $\{b : \phi(b) \leq t\}$ of a quasi star-shaped function are all star-shaped. Constant factors of the following form play a crucial role in our analysis.

Definition 3—*For* $0 \leq \eta^* \leq 1$ *and any pair of quasi star-shaped functions* $\phi_0(b)$ *and* $\phi(b)$*, define a general invertibility factor (GIF) over the cone (*9*) as follows:*

$$F(\xi, S;\varphi_0, \varphi)=\inf\left\{\frac{\Delta(\beta^*+b, \beta^*)e^{\varphi_0(b)}}{|b_S|_1\varphi(b)}:b \in \mathscr{C}(\xi, S), \varphi_0(b) \leq \eta^*\right\}, \quad (14)$$

where $\Delta(\beta, \beta^*)$ *is as in* (4).

The GIF extends the squared compatibility constant (van de Geer and Bühlmann, 2009) and the weak and sign-restricted cone invertibility factors (Ye and Zhang, 2010) from the linear regression model with $\phi_0(\cdot) = 0$ to the general model (1) and from ℓ_q norms to general $\phi(\cdot)$. They are all closely related to the restricted eigenvalues (Bickel et al., 2009; Koltchinskii, 2009) as we will discuss in Subsection 3.1.

The basic inequality (8) implies that the symmetrized Bregman divergence $\Delta(\beta, \hat{\beta}^*)$ is no greater than a linear function of $|h_S|_1$, where $h = \hat{\beta} - \beta^*$. If $\Delta(\beta, \hat{\beta}^*)$ is no smaller than a linear function of the product $|h_S|_1\phi(h)$, then an upper bound for $\phi(h)$ exists. Since the symmetrized Bregman divergence (4) is approximately quadratic, $\Delta(\beta, \hat{\beta}^*) \approx \langle h, \ddot{\psi}(\beta^*)h\rangle$, in a neighborhood of β^*, this is reasonable when $h = \hat{\beta} - \beta^*$ is not too large and $\ddot{\psi}(\beta^*)$ is invertible in the cone. A suitable factor $e^{\phi 0(b)}$ in (14) forces the computation of this lower bound in a proper neighborhood of β^*.

We first provide a set of general oracle inequalities.

Theorem 4: *Let* $\{z^*_0, z^*_1\}$ *be as in (*5*) with* $S \supseteq \{j:\beta^*_j \neq 0\}$*,* Ω_0 *in (*6*),* $0 \leq \eta \leq \eta^* \leq 1$*, and* $\{\phi_0(b), \phi(b)\}$ *be a pair of quasi star-shaped functions. Then, in the event*

$$\Omega_1=\Omega_0 \cap \left\{\frac{|w_S|_\infty\lambda+z^*_0}{(\lambda-z^*_1)_+} \leq \xi, \frac{|w_S|_\infty\lambda+z^*_0}{F(\xi, S;\varphi_0, \varphi_0)} \leq \eta e^{-\eta}\right\}, \quad (15)$$

the following oracle inequalities hold:

$$\varphi_0(\hat{\beta}-\beta^*) \le \eta, \quad \varphi(\hat{\beta}-\beta^*) \le \frac{e^{\eta}(|w_S|_\infty \lambda + z_0^*)}{F(\xi, S; \varphi_0, \varphi)}, \quad (16)$$

and with $\phi_{1,S}(b) = |b_S|_1/|S|$

$$\Delta(\hat{\beta}, \beta^*) + (\lambda - z_1^*)|W_{S^c}(\hat{\beta}-\beta^*)_{S^c}|_1 \le \frac{e^{\eta}(|w_S|_\infty \lambda + z_0^*)^2 |S|}{F(\xi, S; \varphi_0, \varphi_{1,S})}. \quad (17)$$

Remark 5: Sufficient conditions are given in Subsection 3.2 for (15) to hold with high probability. See Lemma 8, Remarks 10 and 11 and Examples 7, 8, and 9.

The oracle inequalities in Theorem 4 control both the estimation error in terms of $\phi(\hat{\beta} - \beta^*)$ and the prediction error in terms of the symmetrized Bregman divergence $\Delta(\beta, \beta^*)$ discussed in Section 2. Since they are based on the GIF (14) in the intersection of the cone and the unit ball $\{b : \phi_0(b) \le 1/e\}$, they are different from typical results in a small-ball analysis based on the Taylor expansion of $\psi(\beta)$ at $\beta = \beta^*$. An important feature of Theorem 4 is that its regularity condition is imposed only on the GIF (14) evaluated at the target β^*; The uniformity of the order of $\Delta(\beta + b, \beta)$ in β is not required. Theorem 4 does allow $\phi_0(\cdot) = 0$ with $F(\xi, S; \phi_0, \phi_0) = \infty$ and $\eta = 0$ in linear regression.

3.1 The Hessian and Related Quantities

In this subsection we describe the relationship between the GIF (14) and the Hessian of the convex function $\psi(\cdot)$ in (1) and examine cases where the quasi star-shaped functions $\phi_0(\cdot)$ and $\phi(\cdot)$ are familiar seminorms. Throughout, we assume that $\psi(\beta)$ is twice differentiable. Let $\ddot{\psi}(\beta)$ be the Hessian of $\psi(\beta)$ and $\Sigma^* = \ddot{\psi}(\beta^*)$.

The GIF (14) can be simplified under the following condition.

Definition 6—*Given a nonnegative-definite matrix* Σ *and constant* $\eta^* > 0$, *the symmetriized Bregman divergence* $\Delta(\beta, \beta^*)$ *satisfies the* ϕ_0*-relaxed convexity* (ϕ_0*-RC) condition if*

$$\Delta(\beta^* + b, \beta^*) e^{\varphi_0(b)} \ge \langle b, \sum b \rangle, \forall b \in \mathscr{C}(\xi, S), \varphi_0(b) \le \eta^*. \quad (18)$$

The ϕ_0-RC condition is related to the restricted strong convexity condition for the Bregman divergence (Negahban et al., 2010): $\ell(\beta^* + b) - \ell(\beta^*) - \langle \dot{\ell}(\beta^*), b \rangle \ge \kappa \|b\|^2$ with a certain restriction $b \in \mathscr{S}$ and a loss function $\|b\|$. It actually implies the restricted strong convexity of the symmetrized Bregman divergence with $\kappa = e^{-\eta^*}$ and loss $\|b\|_* = \langle b, \Sigma b \rangle^{1/2}$. However, (18) is used in our analysis mainly to find a quadratic form as a media for the eventual comparison of $\Delta(\beta^* + b, \beta^*)$ with $|b_S|_1 \phi(b)$ in (14), where $\phi(b)$ is the loss function. In fact, in our examples, we find quasi star-shaped functions ϕ_0 for which (18) holds for unrestricted $b(\eta^* = \xi = \infty)$. In such cases, the ϕ_0-RC condition is a smoothness condition on the Hessian operator $\ddot{\psi}(\beta) = \ddot{\ell}(\beta)$, since $\Delta(\beta^* + h, \beta^*) = \int_0^1 \langle h, \ddot{\psi}(\beta^* + th) h \rangle dt$ by (4).

In what follows, $\Sigma = \Sigma^* = \ddot{\psi}(\beta^*)$ is allowed in all statements unless otherwise stated. Under the ϕ_0-RC (18), the GIF (14) is bounded from below by the following simple GIF:

$$F_0(\xi, S;\varphi)=\inf_{b\in\mathscr{C}(\xi,S)}\frac{\langle b,\sum b\rangle}{|b_S|_1\varphi(b)}. \quad (19)$$

In linear regression, $F_0(\xi, S;\phi)$ is the square of the compatibility factor for $\phi(b) = \phi_{1,S}(b) = |b_S|_1/|S|$ (van de Geer, 2007) and the weak cone invertibility factor for $\phi(b) = \phi_q(b) = |b|_q/|S|^{1/q}$ (Ye and Zhang, 2010). They are both closely related to the restricted isometry property (RIP) (Candes and Tao, 2005), the sparse Rieze condition (SRC) (Zhang and Huang, 2008), and the restricted eigenvalue (Bickel et al., 2009). Extensive discussion of these quantities can be found in Bickel et al. (2009), van de Geer and Bühlmann (2009) and Ye and Zhang (2010). The following corollary is an extension of an oracle inequality of Ye and Zhang (2010) from linear regression to the general convex minimization problem (1).

Corollary 7: *Let* $\eta \le \eta^* \le 1$. *Suppose the* ϕ_0*-RC condition (*18*). Then, in the event*

$$\Omega_0 \cap \{|w_S|_\infty\lambda+z_0^* \le \min(\xi(\lambda-z_1^*), \eta e^{-\eta}F_0(\xi, S;\varphi_0))\},$$

*the oracle inequalities (*16*) and (*17*) in Theorem 4 hold with the GIF* $F(\xi, S;\phi_0, \phi)$ *replaced by the simple GIF* $F_0(\xi, S; \phi)$ *in (*19*). In particular, in the same event,*

$$\varphi_0(h) \le \eta, |h|_q \le \frac{e^\eta(|w_S|_\infty\lambda+z_0^*)|S|^{1/q}}{F_0(\xi, S;\varphi_q)}, \forall q>0,$$

with $\phi_q(b) = |b|_q/|S|^{1/q}$ *and* $h = \hat{\beta} - \beta^*$, *and with* $\phi_{1,S}(b) = |b_S|_1/|S|$,

$$e^{-\eta}\langle h, \sum h\rangle \le \Delta(\hat{\beta}, \beta^*) \le \frac{e^\eta(|w_S|_\infty\lambda+z_0^*)^2|S|}{F_0(\xi, S;\varphi_{1,S})}-(\lambda-z_1^*)|W_{S^c}h_{S^c}|_1.$$

Here the only differences between the general model (1) and linear regression ($\phi_0(b) = 0$) are the extra factor e^η with $\eta \le 1$, the extra constraint $|w_S|_\infty\lambda+z_0^* \le \eta e^{-\eta}F_0(\xi, S;\varphi_0)$, and the extra ϕ_0-RC condition (18). Moreover, the simple GIF (19) explicitly expresses all conditions on $F_0(\xi, S;\phi)$ as properties of a fixed matrix Σ.

Example 6 (Linear regression: oracle inequalities): *For* $\psi(\beta)=|Xb|_2^2/(2n)$ *and* $\Sigma = X'X/n$, $F_0(\xi, S;\phi_q)$ *is the weak cone invertibility factor for* $q \in [1,\infty]$ (Ye and Zhang, 2010), *where a sharper version is defined as the sign restricted invertibility factor (SCIF):*

$$\mathrm{SCIF}_q(\xi, S)=\inf_{b\in\mathscr{C}_-(\xi,S)}|\sum b|_\infty/\varphi_q(b), \varphi_q=|b|_q/|S|^{1/q}.$$

For $q = 1$, $F_0^{1/2}(\xi, S;\varphi_{1,S})$ *is the compatibility constant (*van de Geer, 2007*)*

$$\kappa_*(\xi,S)=\inf_{b\in\mathscr{C}(\xi,S)}\frac{|S|^{1/2}|Xb|_2}{|b_S|_1 n^{1/2}}=\inf_{b\in\mathscr{C}(\xi,S)}\left(\frac{b'\sum b}{|b_S|_1^2/|S|}\right)^{1/2}. \quad (20)$$

They are all closely related to the ℓ_2 restricted eigenvalues

$$\mathrm{RE}_2(\xi,S)=\inf_{b\in\mathscr{C}(\xi,S)}\frac{|Xb|_2}{|b|_2 n^{1/2}}=\inf_{b\in\mathscr{C}(\xi,S)}\left(\frac{b'\sum b}{|b|_2^2}\right)^{1/2}$$

(Bickel et al., 2009; Koltchinskii, 2009). *Since* $|b_S|_1^2 \le |b|_2^2|S|$, $\kappa_*(\xi,S) \ge \mathrm{RE}_2(\xi, S)$ (van de Geer and Bühlmann, 2009). *For the Lasso with* $\hat{w}_j = 1$,

$$|\widehat{\beta}-\beta^*|_2 \le \frac{|S|^{1/2}(\lambda+z_0^*)}{\mathrm{SCIF}_2(\xi,S)} \le \frac{|S|^{1/2}(\lambda+z_0^*)}{F_0(\xi,S;\varphi_2)} \le \frac{|S|^{1/2}(\lambda+z_0^*)}{\kappa_*(\xi,S)\mathrm{RE}_2(\xi,S)} \quad (21)$$

in the event $\lambda+z_0^* \le \xi(\lambda-z_1^*)$ (Ye and Zhang, 2010). *Thus, cone and general invertibility factors yield sharper ℓ_2 oracle inequalities.*

The factors in the oracle inequalities in (21) do not always have the same order for large $|S|$. Although the oracle inequality based on $\mathrm{SCIF}_2(\xi, S)$ is the sharpest among them, it seems not to lead to a simple extension to the general convex minimization in (1). Thus, we settle with extensions of the second sharpest oracle inequality in (21) with $F_0(\xi, S;\cdot)$.

3.2 Oracle Inequalities for the Lasso in GLM

An important special case of the general formulation is the ℓ_1-penalized estimator in a generalized linear model (GLM) (McCullagh and Nelder, 1989). This is Example 3 in Subsection 2.2, where we set up the notation in (13) and gave the KL divergence interpretation to (4). The ℓ_1 penalized, normalized negative likelihood is

$$\ell(\beta)=\psi(\beta)-z'\beta, \text{ with } \psi(\beta)=\frac{1}{n}\sum_{i=1}^n\{\psi_0(x^i\beta)-c(y_i,\sigma)\} \text{ and } z=\frac{X'y}{n}. \quad (22)$$

Assume that ψ_0 is twice differentiable. Denote the first and second derivatives of ψ_0 by $\dot{\psi}_0$ and $\ddot{\psi}_0$, respectively. The gradient and Hessian are

$$\dot{\psi}(\beta)=X'\psi_0(\theta)/n \text{ and } \ddot{\psi}(\beta)=X'\mathrm{diag}(\ddot{\psi}_0(\theta))X/n, \quad (23)$$

where $\theta = X\beta$ and $\dot{\psi}_0$ and $\ddot{\psi}_0$ are applied to the individual components of θ.

A crucial condition in our analysis of the Lasso in GLM is the Lipschitz condition

$$\max_{i\le n}\left|\log(\ddot{\psi}_0(x^i\beta^*+t))-\log(\ddot{\psi}_0(x^i\beta^*))\right| \le M_1|t|, \forall M_1|t| \le \eta^*, \quad (24)$$

where M_1 and η^* are constants determined by ψ_0. This condition gives

$$\Delta(\beta^*+b,\beta^*)=\int_0^1\langle b,\ddot{\psi}(\beta^*+tb)b\rangle dt \geq \int_0^1 \sum_{tM_1|x^ib|\leq\eta^*} \frac{\ddot{\psi}_0(x^i\beta^*)\,(x^ib)^2}{ne^{tM_1|x^ib|}}dt,$$

which implies the following lower bound for the GIF in (14):

$$F(\xi,S;\varphi_0,\varphi) \geq \inf_{b\in\mathscr{C}(\xi,S),\varphi_0(b)\leq\eta^*}\sum_{i=1}^n \frac{\ddot{\psi}_0(x^i\beta^*)\,(x^ib)^2}{n|b_S|_1\varphi(b)}\int_0^1 I\{tM_1|x^ib|\ \leq\varphi_0(b)\}dt.$$

For seminorms ϕ_0 and ϕ, the infimum above can be taken over $\phi_0(b) = M_2$ due to scale invariance. Thus, for $\phi_0(b) = M_2|b|_2$ and seminorms ϕ, this lower bound is

$$F^*(\xi,S;\varphi)=\inf_{b\in\mathscr{C}(\xi,S),|b|_2=1}\sum_{i=1}^n \frac{M_2\ddot{\psi}_0(x^i\beta^*)}{n|b_S|_1\varphi(b)}\min\left(\frac{|x^ib|}{M_1},\frac{(x^ib)^2}{M_2}\right), \quad (25)$$

due to $(x^ib)^2\int_0^1 I\{tM_1|x^ib|\ \leq M_2\}dt=M_2\min\{|x^ib|/M_1,(x^ib)^2/M_2\}$.

If (24) holds with $\eta^* = \infty$, $\Delta(\beta^*+b,\beta^*)\geq n^{-1}\int_0^1\sum_i\ddot{\psi}_0(x^i\beta^*)\,(x^ib)^2e^{-tM_1|x^ib|}dt$, so that by the Jensen inequality (18) holds with $\Sigma = \Sigma^* = \psi(\ddot{\beta}^*)$ and

$$\varphi_0(b)=\frac{M_1\sum_{i=1}^n\ddot{\psi}_0(x^i\beta^*)|x^ib|^3}{\sum_{i=1}^n\ddot{\psi}_0(x^i\beta^*)\,(x^ib)^2}\leq M_1|Xb|_\infty. \quad (26)$$

This gives a special $F_0(\xi, S;\phi_0)$ as

$$F_*(\xi,S)=\inf_{b\in\mathscr{C}(\xi,S)}\frac{n\langle b,\sum^*b\rangle^2/(M_1|b_S|_1)}{\sum_{i=1}^n\ddot{\psi}_0(x^i\beta^*)|x^ib|^3}. \quad (27)$$

Since $|Xb|_\infty \leq |X_S|_\infty|b_S|_1+|X_{S^c}W_{S^c}^{-1}|_\infty|W_{S^c}b_{S^c}| \leq\{|X_S|_\infty+\xi|X_{S^c}W_{S^c}^{-1}|_\infty\}|b_S|$ in the cone $\mathscr{C}(\xi,S)$ in (9), for $\phi_0(b) = M_3|b_S|_1$ with $M_3=M_1\{|X_S|_\infty+\xi|X_{S^c}W_{S^c}^{-1}|_\infty\}$, the ϕ_0-RC condition (18) automatically implies the stronger

$$e^{-\varphi_0(b)}\langle b,\sum{}^*b\rangle \leq \Delta(\beta^*+b,\beta^*)\leq e^{\varphi_0(b)}\langle b,\sum{}^*b\rangle,\forall b\in\mathscr{C}(\xi,S),\ \varphi_0(b)<\infty. \quad (28)$$

Under the Lipschitz condition (24), we may also use the following large deviation inequalities to find explicit penalty levels to guarantee the noise bound (15).

Lemma 8

i. *Suppose the model conditions (*13*) and (*24*) with certain* $\{M_1, \eta^*\}$. *Let* x_j *be the columns of X,* $\sum_{ij}^*$ *be the elements of* $\Sigma^* = \psi(\ddot{\beta}^*)$. *For penalty levels* $\{\lambda_0, \lambda_1\}$ *define* $t_j = \lambda_0 I\{j \in S\} + w_j\lambda_1 I\{j \notin S\}$. *Suppose the bounds* w_j *in (*6*) are deterministic and*

$$M_1 \max_{j \le p} (|x_j|_\infty |t_j / \sum\nolimits_{jj}^*) \le \eta_0 e^{\eta_0} \text{ and } \sum_{j=1}^{p} \exp\left\{ -\frac{n t_j^2 e^{-\eta_0}}{2\sigma^2 \sum_{jj}^*} \right\} \le \frac{\varepsilon_0}{2} \quad (29)$$

for certain constants $\eta_0 \le \eta^*$ *and* $\varepsilon_0 > 0$. *Then,* $\mathrm{P}_{\beta^*}\{z_0^* \le \lambda_0, z_1^* \le \lambda_1\} \ge 1 - \varepsilon_0$.

ii. *If* $c_0 = \max_t \ddot{\psi}(t)$, *then part (i) is still valid if (24) and (29) are replaced by*

$$\sum_{j=1}^{p} \exp\left\{ -\frac{n^2 t_j^2}{2\sigma^2 c_0 |x_j|_2^2} \right\} \le \frac{\varepsilon_0}{2}. \quad (30)$$

In particular, if $|x_j|_2^2 = n$, $1 \le j \le p$, $w_j = 1$, $j \notin S$ *and* $\lambda_0 = \lambda_1 = \lambda$ (*so* $t_j = \lambda$), *then part (i) still holds if* $\lambda \ge \sigma \sqrt{(2c_0/n)\log(2p/\varepsilon_0)}$.

The following theorem is a consequence of Theorem 4, Corollary 7 and Lemma 8.

Theorem 9

i. *Let* $\hat{\beta}$ *be the weighted Lasso estimator in (2) with GLM loss function in (22). Let* β^* *be a target vector and* $h = \hat{\beta} - \beta^*$. *Suppose that the data follows the GML model (13) satisfying the Lipschitz condition (24) with certain* $\{M_1, \eta^*\}$. *Let* $F^*(\xi, S; \varphi)$ *be as in (25) with* $S \supseteq \{j : \beta_j^* \ne 0\}$ *and a constant* M_2. *Let* $\eta \le 1 \wedge \eta^*$ *and* $\{\lambda, \lambda_0, \lambda_1\}$ *satisfy*

$$|w_S|_\infty \lambda + \lambda_0 \le \min\{\xi(\lambda - \lambda_1), \eta e^{-\eta} F^*(\xi, S; M_2 |\cdot|_2)\}. \quad (31)$$

Then, in the event $\Omega_0 \cap \{\max_{k=0,1}(z_k^*/\lambda_k) \le 1\}$ *with the* z_k^* *in (5) and* Ω_0 *in (6),*

$$\Delta(\beta^* + h, \beta^*) \le \frac{e^\eta (|w_S|_\infty \lambda + \lambda_0)^2 |S|}{F^*(\xi, S; \varphi_{1,S})}, \quad \varphi(h) \frac{e^\eta (|w_S|_\infty \lambda + \lambda_0)}{F^*(\xi, S; \varphi)} \quad (32)$$

for any seminorm φ *as the estimation loss. In particular, for* $\varphi(b) = M_2 |b|_2$, *(32) gives* $|h|_2 \le \eta / M_2$. *Moreover, if either (29) or (30) holds for the penalty level* $\{\lambda_0, \lambda_1\}$ *and the weight bounds* w_j *in (6) are deterministic, then*

$$\mathrm{P}_{\beta^*}\{(32) \text{ holds for all seminorms } \varphi\} \ge \mathrm{P}_{\beta^*}(\Omega_0) - \varepsilon_0.$$

ii. *Suppose* $\eta^* = \infty$ *and (31) holds with* $F^*(\xi, S; M_2 |\cdot|_2)$ *replaced by the special simple GIF* $F_*(\xi, S)$ *in (27) for the* φ_0 *in (26). Then, the conclusions of part (i) hold with* $F^*(\xi, S; \cdot)$ *replaced by the simple GIF* $F_0(\xi, S; \cdot)$ *in (19). Moreover,* $\varphi_0(h) \le \eta$ *and (32) can be strengthened with the lower bound* $\Delta(\beta^* + h, \beta^*) \ge e^{-\eta} \langle h, \Sigma^* h \rangle$.

iii. *For any* $\eta^* > 0$, *the conclusions of part (ii) hold for the* $\varphi_0(b) = M_3 |b_S|_1$ *in (28), if* $F_*(\xi, S)$ *is replaced by* $\kappa_*^2(\xi, S)/(M_3 |S|)$ *in (31), where* $\kappa_*(\xi, S)$ *is the compatibility constant in (20).*

Remark 10: *If either (29) or (30) holds for the penalty levels* $\{\lambda_0, \lambda_1\}$ *and the bounds* w_j *in (6) are deterministic, then (32) implies* $\mathrm{P}_{\beta^*}\{$*the noise bound (15) holds*$\} \ge \mathrm{P}_{\beta^*}(\Omega_0) - \varepsilon_0$.

Remark 11: *Suppose that* $\max_{j\notin S} 1/w_j$, $\max_j 1/\sum^*_{jj}$, $\max_{j\in S} w_j$, $\max_j \sum^*_{jj}$, *and* M_1 *are all bounded, and that* $\{1+F_*^2(\xi,S)\}(\log p)/n \to 0$. *Then, (29) holds with the penalty level* $\lambda_0=\lambda_1=a\sigma\sqrt{(2/n)\log(p/\varepsilon_0)}$ *for certain* $a \le (1+o(1))\max_j(\sum^*)^{1/2}_{jj}/w_j$, *due to* $\max\{\lambda_0, \eta, \eta_0\} \to 0+$. *Again, the conditions and conclusions of Theorem 9 "converge" to those for the linear regression as if the Gram matrix is* Σ^*.

Remark 12: *In Theorem 9, the key condition (31) is weaker in parts (i) and (ii) than part (iii), although part (ii) requires* $\eta^*=\infty$. *For* $\Sigma=\Sigma^*$ *and* $M_1=M_2\le M_3/(1+\xi)$,

$$\kappa_*^2(\xi,S)/(M_3|S|) \le \min\left\{F_*(\xi,S), F^*(\xi,S;M_2|\cdot|_2)\right\},$$

since $n^{-1}\sum_{i=1}^n \ddot\psi_0(x^i\beta^*)|x^ib|^3/\langle b,\sum^* b\rangle \le |Xb|_\infty \le |b_S|_1 M_3/M_1$ *as in the derivation of (28) and* $|b|_2\le|b|_1\le(1+\xi)|b_S|_1$ *in the cone (9). For the more familiar* $\kappa_*^2(\xi,S)/(M_3|S|)$ *with the compatibility constant, (31) essentially requires a small* $|S|\sqrt{(\log p)/n}$. *The sharper Theorem 9 (i) and (ii) provides conditions to relax the requirement to a small* $|S|(\log p)/n$.

Remark 13: *For* $\hat w_j = 1$, Negahban et al. (2010) *considered M-estimators under the restricted strong convexity condition discussed below Definition 6. For the GLM, they considered iid sub-Gaussian* x^i *and used empirical process theory to bound the ratio* $\Delta(\beta^*+b,\beta^*)/\{|b|_2(|b|_2-c_0|b|_1\}$ *from below over the cone (9) with a small* c_0. *Their result extends the* ℓ_2 *error bound* $|S|^{1/2}(\lambda+z_0^*)/\mathrm{RE}_2^2(\xi,S)$ *of* Bickel et al. (2009), *while Theorem 9 extends the sharper (21) with the factor* $F_0(\xi,S;\phi_2)$. *Theorem 9 applies to both deterministic and random designs. Similar to* Negahban et al. (2010), *for iid sub-Gaussian* x^i, *empirical process theory can be applied to the lower bound (25) for the GIF to verify the key condition (31) with* $F^*(\xi,S;M_2|\cdot|_2) \gtrsim |S|^{-1/2}$, *provided that* $|S|(\log p)/n$ *is small.*

Example 7 (Linear regression: oracle inequalities, continuation): *For the linear regression model (10) with quadratic loss,* $\psi_0(\theta)=\theta^2/2$, *so that (24) holds with* $M_1=0$ *and* $\eta^*=\infty$. *It follows that* $F^*(\xi,S;M_2|\cdot|_2)=\infty$ *and (31) has the interpretation with* $\eta=0+$ *and* $\eta e^{-\eta}F^*(\xi,S;M_2|\cdot|_2)=\infty$. *Moreover, since* $M_1=0$, $\eta_0=0+$ *in (29). Thus, the conditions and conclusions of Theorem 9 "converge" to the case of linear regression as* $M_1\to 0+$. *Suppose iid* $\varepsilon_i \sim N(0,\sigma^2)$ *as in (13). For* $\hat w_j = w_j = 1$ *and* $\sum^*_{jj}=\sum_{i=1}^n x_{ij}^2/n=1$, *(29) holds with* $\lambda_0=\lambda_1=\sigma\sqrt{(2/n)\log(p/\varepsilon_0)}$ *and (31) holds with* $\lambda=\lambda_0(1+\xi)/(1-\xi)$. *The value of* σ *can be estimated iteratively using the mean residual squares (*Städler et al., 2010*;* Sun and Zhang, 2011*). Alternatively, cross-validation can be used to pick* λ. *For* $\phi(b)=\phi_2(b)=|b|_2/|S|^{1/2}$, *(32) matches the risk bound in (21) with the factor* $F_0(\xi,S;\phi_2)$.

Example 8 (Logistic regression: oracle inequalities): *The model and loss function are given in (11) and (12) respectively. Here we verify the conditions of Theorem 9. The Lipschitz condition (24) holds with* $M_1=1$ *and* $\eta^*=\infty$ *since* $\psi_0(t)=\log(1+e^t)$ *provides*

$$\frac{\ddot\psi_0(\theta+t)}{\ddot\psi_0(\theta)}=\frac{e^t(1+e^\theta)^2}{(1+e^{\theta+t})^2}\ge\begin{cases} e^{-|t|} & t<0\\ e^{-t}(1+e^\theta)^2/(e^{-t}+e^\theta)^2 \ge e^{-|t|} & t>0.\end{cases}$$

Since $\max_t \ddot{\psi}(t) = c_0 = 1/4$ *we can apply (*30*). In particular, if* $\widehat{w}_j = w_j = 1 = |x_j|_2^2/n$, $\lambda = \{(\xi+1)/(\xi-1)\}\sqrt{(\log(p/\varepsilon_0))/(2n)}$ *and* $\lambda\{2\xi/(\xi+1)\}/F_*(\xi,S) \le \eta e^{-\eta}$*, then (*32*) holds with at least probability* $1-\varepsilon_0$ *under* $\mathrm{P}_\beta{}^*$*. For such deterministic* $\hat{W}$ *and X, an adaptive choice of the penalty level is* $\lambda = \widehat{\sigma}\sqrt{(2/n)\log p}$ *with* $\widehat{\sigma}^2 = \sum_{i=1}^n \pi_i(\widehat{\beta})\{1-\pi_i(\widehat{\beta})\}/n$*, where* $\pi_i(\beta)$ *is as in Example 2.*

Example 9 (Log-linear models: oracle inequalities): *Consider counting data with* $y_i \in \{0,1,2,\ldots\}$*. In log-linear models, it is assume that*

$$\mathrm{E}_\beta(y_i) = e^{\theta_i}, \theta_i = x^i\beta,\ 1 \le i \le n.$$

This becomes a GLM with the average negative Poisson log-likelihood function

$$\ell(\beta) = \psi(\beta) - z'\beta,\ \psi(\beta) = \sum_{i=1}^n \frac{\exp(x^i\beta) - \log(y_i!)}{n},\ z = X'y/n.$$

In this model, $\psi_0(t) = e^t$*, so that the Lipschitz condition (*24*) holds with* $M_1 = 1$ *and* $\eta^* = \infty$*. Although (*30*) is not useful with* $c_0 = \infty$*, (*29*) can be used in Theorem 9.*

4. Adaptive and Multistage Methods

We consider in this section an adaptive Lasso and its repeated applications, with weights recursively generated from a concave penalty function. This approach appears to provide the most appealing choice of weights both from heuristic and theoretical standpoints. The analysis here uses the results in Section 3 and an idea in Zhang (2010b).

We first consider adaptive Lasso and provide conditions under which it improves upon its initial estimator. Let $\rho_\lambda(t)$ be a concave penalty function with $\rho_\lambda(0+) = \lambda$, where $\dot{\rho}_\lambda(t) = (\partial/\partial t)\rho_\lambda(t)$. The maximum concavity of the penalty is

$$\kappa = \sup_{0<t_1<t_2} \frac{|\dot{\rho}_\lambda(t_2) - \dot{\rho}_\lambda(t_1)|}{t_2 - t_1}. \quad (33)$$

Let $\mathscr{C}(\xi,S)$ be the cone in (9). Let $\phi_0(b)$ be a quasi star-shaped function and define

$$F_2(\xi, S; \varphi_0) = \inf\left\{\frac{e^{\varphi_0(b)}\Delta(\beta^* + b, \beta^*)}{|b_S|_2|b|_2} : 0 \ne b \in \mathscr{C}(\xi,S), \varphi_0(b) \le \eta^*\right\}. \quad (34)$$

This quantity is an ℓ_2 version of the GIF in (14). The analysis in Section 3 can be used to find lower bounds for (34) in the same way simply by taking $\phi(b) = |b|_2$ and replacing $|b_S|_1$ with $|b_S|_2$. For example, in generalized linear models (13) satisfying the Lipschitz condition (24), the derivation of (25) yields

$$F_2(\xi, S; M|\cdot|_2) \geq \inf_{b\in\mathscr{C}(\xi,S),|b|_2=1} \sum_{i=1}^{n} \frac{M_2\ddot{\psi}_0(x^i\beta^*)}{n|b_S|_2} \min\left(\frac{|x^ib|}{M_1}, \frac{(x^ib)^2}{M_2}\right).$$

Given $0 < \varepsilon_0 < 1$, the components of the error vector $z - \dot{\psi}(\beta^*)$ are sub-Gaussian if for all $0 \leq t \leq \sigma\sqrt{(2/n)\log(4p/\varepsilon_0)}$,

$$\mathrm{P}_{\beta^*}\left\{|(z-\dot{\psi}(\beta^*))_j| \geq t\right\} \leq 2e^{-nt^2/(2\sigma^2)}. \quad (35)$$

This condition holds for all GLM when the components of $X\beta^*$ are uniformly in the interior of the natural parameter space for the exponential family.

Theorem 14

Let κ be as in (33), $S_0=\{j:\beta_j^* \neq 0\}$, $\lambda_0 > 0$, $0 < \eta < 1$, $0 < \gamma_0 < 1/\kappa$, $A > 1$, and $\xi \geq (A+1-\kappa\gamma_0)/(A-1)$. Let ϕ_0 be a quasi star-shaped function, $F(\xi, S;\phi_0, \phi_0)$ be the GIF in (14), and $F_2(\xi, S;\phi_0)$ its ℓ_2-version in (34). Suppose

$$\lambda_0\{1+A/(1-\kappa\gamma_0)\} \leq F(\xi, S;\varphi_0, \varphi_0)\eta e^{-\eta}, F_* \leq F_2(\xi, S;\varphi_0), \quad (36)$$

for all $S \supseteq S_0$ with $|S\backslash S_0| \leq \ell^*$. Let $\tilde{\beta}$ be an initial estimator of β and $\hat{\beta}$ be the weighted Lasso in (2) with weights $\hat{w}_j = \rho_\lambda(|\tilde{\beta}_j|)/\lambda$ and penalty level $\lambda = A\lambda_0/(1-\kappa\gamma_0)$. Then,

$$|\hat{\beta}-\beta^*|_2 \leq \frac{e^\eta}{F_*}\left\{|\dot{\rho}_\lambda(|\beta^*_{S_0}|)|_2 + |\{z-\dot{\psi}(\beta^*)\}_{S_0}|_2 + \left(\kappa+\frac{1}{\gamma_0 A}-\frac{\kappa}{A}\right)|\tilde{\beta}-\beta^*|_2\right\}$$

in the event $\{|(\tilde{\beta}-\beta)_{S_0^c}|_2^2 \leq \gamma_0^2\lambda^2\ell^*\} \cap \{|z-\dot{\psi}(\beta^*)|_\infty \leq \lambda_0\}$. Moreover, if (35) holds and $\lambda_0 = \sigma\sqrt{(2/n)\log(2p/\varepsilon_0)}$ with $0 < \varepsilon_0 < 1$, then $\mathrm{P}_{\beta^*}\{|z-\ddot{\psi}(\beta^*)| \geq \lambda_0\} \leq \varepsilon_0$.

Theorem 14 raises the possibility that $\hat{\beta}$ improves $\tilde{\beta}$ under proper conditions. Thus it is desirable to repeatedly apply this adaptive Lasso in the following way,

$$\hat{\beta}^{(k+1)} = \arg\min_\beta \left\{\ell(\beta) + \sum_{j=1}^{p} \dot{\rho}_\lambda(\hat{\beta}_j^{(k)})|\beta_j|\right\}, k = 0, 1, \ldots. \quad (37)$$

Such multistage algorithms have been considered in the literature (Fan and Li, 2001; Zou and Li, 2008; Zhang, 2010b). As discussed in Remark 16 below, it is beneficial to use a concave penalty ρ_λ in (37). Natural choices of ρ_λ include the smoothly clipped absolute deviation and minimax concave penalties (Fan and Li, 2001; Zhang, 2010a).

Theorem 15

Let $\{\kappa, S_0, \lambda_0, \eta, \gamma_0, A, \xi, \ell^*, \lambda\}$ be the same as Theorem 14. Let $\hat{\beta}^{(0)}$ be the unweighted Lasso with $\hat{w}_j = 1$ in (2) and $\hat{\beta}^{(\ell)}$ be the ℓ-th iteration of the recursion (37) initialized with $\hat{\beta}^{(0)}$. Let $\xi_0 = (\lambda + \lambda_0)/(\lambda - \lambda_0)$. Suppose (36) holds and

$$e^{\eta}\{1+(1-\kappa\gamma_0)/A\}/F(\xi_0, S_0;\varphi_0, |\cdot|_2) \le \gamma_0\sqrt{\ell^*}. \quad (38)$$

Define $r_0 = (e^{\eta}/F_*)\{\kappa + 1/(\gamma_0 A) - \kappa/A\}$. *Suppose* $r_0 < 1$. *Then,*

$$|\widehat{\beta}^{(\ell)}-\beta^*|_2 \le \frac{|\dot\rho_\lambda(|\beta^*_{S_0}|)|_2+|\{z-\dot\psi(\beta^*)\}_{S_0}|_2}{e^{-\eta}F_*(1-r_0)/(1-r_0^\ell)}+\frac{r_0^\ell e^{\eta}\lambda\{1+(1-\kappa\gamma_0)/A\}}{F(\xi_0, S_0;\varphi_0, |\cdot|_2)} \quad (39)$$

in the event

$$\left\{|z-\dot\psi(\beta^*)|_\infty \le \lambda_0\right\} \cap \left\{\frac{|\dot\rho_\lambda(|\beta^*_{S_0}|)|_2+|\{z-\dot\psi(\beta^*)\}_{S_0}|_2}{e^{-\eta}F_*(1-r_0)} \le \gamma_0\lambda\sqrt{\ell^*}\right\}. \quad (40)$$

Moreover, if (35) *holds and* $\lambda_0=\sigma\sqrt{(2/n)\log(4p/\varepsilon_0)}$ *with* $0 < \varepsilon_0 < 1$, *then the intersection of the events* (40) *and* $\{|\{z-\dot\psi(\beta^*)\}_{S_0}|_2 \le n^{-1/2}\sigma\sqrt{2|S_0|\log(4|S_0/\varepsilon_0|)}\}$ *happens with at least* P_{β^*} *probability* $1 - \varepsilon_0$, *provided that*

$$\frac{|\dot\rho_\lambda(|\beta^*_{S_0}|)|_2+n^{-1/2}\sigma\sqrt{2|S_0|\log(4|S_0|/\varepsilon_0)}}{e^{-\eta}F_*(1-r_0)} \le \frac{\gamma_0 A\lambda_0\sqrt{\ell^*}}{1-\kappa\gamma_0}.$$

Remark 16

Define $R^{(0)} = \lambda e^{\eta}\{1 + (1 - \kappa\gamma_0)/A\}/F(\xi_0, S_0; \phi_0, |\cdot|_2)$ *and*

$$R^{(\infty)}=\frac{|\dot\rho_\lambda(|\beta^*_{S_0}|)|_2+|\{z-\dot\psi(\beta^*)\}_{S_0}|_2}{e^{-\eta}F_*(1-r_0)}, R^{(\ell)}=(1-r_0^\ell)R^{(\infty)}+r_0^\ell R^{(0)}.$$

It follows from (39) *that* $R^{(\ell)}$ *is an upper bound of* $|\hat\beta^{(\ell)} - \beta^*|_2$ *under proper conditions. This implies* $|\hat\beta^{(\ell)} - \beta^*| \le 2R^{(\infty)}$ *after* $\ell = |\log r_0|^{-1}\log(R^{(\infty)}/R^{(0)})$ *iterations of the recursion* (37). *Under condition* (35),

$$E_{\beta^*}R^{(\infty)} \le \{|\dot\rho_\lambda(|\beta^*_{S_0}|)|_2+2\sigma\sqrt{|S_0|/n}\}e^{\eta}/\{F_*(1-r_0)\}.$$

Since $\rho_\lambda(t)$ *is concave in* t, $|\dot\rho_\lambda(|\beta^*_{S_0}|)|_2 \le \dot\rho_\lambda(0+)|S_0|^{1/2}=\lambda|S_0|^{1/2}$. *This component of* $E_{\beta^*}R^{(\infty)}$ *matches the noise inflation due to model selection uncertainty since* $\lambda \asymp \lambda_0=\sigma\sqrt{(2/n)\log(p/\varepsilon_0)}$. *This noise inflation diminishes when* $\min_{j\in S_0}|\beta_j^*| \ge \gamma\lambda$ *and* $\dot\rho_\lambda(t) = 0$ *for* $|t| \ge \gamma\lambda$, *yielding the super-efficient* $E_{\beta^*}R^{(\infty)} \le \{2\sigma\sqrt{|S_0|/n}\}e^{\eta}/\{F_*(1-r_0)\}$ *without the* log *p factor. The risk bound* $R^{(\infty)}$ *is comparable with those for concave penalized least squares in linear regression* (Zhang, 2010a).

Remark 17

For $\log(p/n) \asymp \log p$*, the penalty level* λ *in Theorems 14 and 15 are comparable with the best proven results and of the smallest possible order in linear regression. For* $\log(p/n) \ll \log p$*, the proper penalty level is expected to be of the order* $\sigma\sqrt{(2/n)\log(p/|S_0|)}$ *under a vectorized sub-Gaussian condition which is slightly stronger than (*35*). This refinement for* $\log(p/n) \ll \log p$ *is beyond the scope of this paper.*

5. Selection Consistency

In this section, we provide a selection consistency theorem for the ℓ_1 penalized convex minimization estimator, including both the weighted and unweighted cases. Let $\|M\|_\infty = \max_{|u|_\infty \le 1} |Mu|_\infty$ be the ℓ_∞-to-ℓ_∞ operator norm of a matrix M.

Theorem 18

Let $\psi(\beta) = \ell(\beta)$ *be the Hessain of the loss in (*1*),* $\hat{\beta}$ *be as in (*2*),* β^* *be a target vector,* z_k^* *be as in (*5*),* Ω_0 *in (*6*),* $S \supseteq \{j:\beta_j^* \neq 0\}$ *and* $F(\xi,S;\phi_0,\phi)$ *as in* (14).

i. *Let* $0 < \eta \le \eta^* \le 1$, $\mathscr{B}_0^* = \{\beta:\varphi_0(\beta-\beta^*) \le \eta, \beta_{S^c}=0\}$ *and* $S_\beta = \{j: \beta_j \neq 0\}$. *Suppose*

$$\sup_{\beta\in\mathscr{B}_0^*} |\widehat{W}_{S^c}^{-1}\ddot{\psi}_{S^c,S_\beta}(\beta)\{\ddot{\psi}_{S_\beta}(\beta)\}^{-1}\widehat{W}_{S_\beta}\mathrm{sgn}(\beta_{S_\beta})|_\infty \le \kappa_0<1 \quad (41)$$

$$\sup_{\beta\in\mathscr{B}_0^*} \|\widehat{W}_{S^c}^{-1}\ddot{\psi}_{S^c,S_\beta}(\beta)\{\ddot{\psi}_{S_\beta}(\beta)\}^{-1}\|_\infty \le \kappa_1. \quad (42)$$

Then, $\{j: \hat{\beta}_j \neq 0\} \subseteq S$ *in the event*

$$\Omega_1^* = \Omega_0 \cap \{|\hat{w}_S|_\infty\lambda + z_0^* \le \eta e^{-\eta}F(0,S;\varphi_0,\varphi_0), \kappa_1 z_0^* + z_1^* < (1-\kappa_0)\lambda\}. \quad (43)$$

ii. *Let* $0 < \eta \le \eta^* \le 1$ *and* $\mathscr{B}_0 = \{\beta: \phi_0(\beta-\beta^*) \le \eta, \mathrm{sgn}(\beta) = \mathrm{sgn}(\beta^*)\}$. *Suppose (*41*) and (*42*) hold with* $\mathscr{B}_0^*$ *replaced by* $\mathscr{B}_0$. *Then,* $\mathrm{sgn}(\hat{\beta}) = \mathrm{sgn}(\beta^*)$ *in the event*

$$\Omega_1^* \cap \left\{\sup_{\beta\in\mathscr{B}_0} \|\{\ddot{\psi}_S(\beta)\}^{-1}\|_\infty(|\hat{w}_S|_\infty\lambda + z_0^*) < \min_{j\in S}|\beta_j^*|\right\}. \quad (44)$$

iii. *Suppose conditions of Theorem 9 hold for the GLM. Then, the conclusions of (i) and (ii) hold under the respective conditions if* $F(0,S;\phi_0,\phi_0)$ is replaced by $F^*(\xi,S;M_2|\cdot|_2)$ *or* $F_*(\xi,S)$ *or* $\kappa_*^2(\xi,S)/(M_3|S|)$ *with the respective* ϕ_0 *in Theorem 9.*

For $\hat{w}_j = 1$, this result is somewhat more specific in the radius η for the uniform irrepresentable condition (41), compared with a similar extension of the selection consistency theory to the graphical Lasso by Ravikumar et al. (2008). In linear regression (10), $\psi(\beta) = \Sigma = X'X/n$ does not depend on β, so that Theorem 18 with the special $w_j = 1$ matches the existing selection consistency theory for the unweighted Lasso (Meinshausen and Bühlmann, 2006; Tropp, 2006; Zhao and Yu, 2006; Wainwright, 2009). We discuss below the ℓ_1 penalized logistic regression as a specific example.

Example 10 (Logistic regression: selection consistency)—*Suppose* $w_j = 1 = |x_j|_2^2/n$ *where* x_j *are the columns of X. If (*43*) and (*44*) hold with* z_0^* *and* z_1^* *replaced*

by $\sqrt{(\log(p/\varepsilon_0))/(2n)}$, *then the respective conclusions of Theorem 18 hold with at least probability* $1 - \varepsilon_0$ *in* P_{β^*}.

6. The Sparsity of the Lasso and SRC

The results in Sections 2, 3, and 4 are concerned with prediction and estimation properties of $\hat{\beta}$, but not dimension reduction. Theorem 18 (i) and (iii) provide dimension reduction under ℓ_∞-type conditions (41) and (42). In this section, we provide upper bounds for the dimension of $\hat{\beta}$ under conditions of a weaker ℓ_2 type. For this purpose, we introduce

$$\kappa_+(m) = \sup_{|B|=m} \left\{ \lambda_{\max}\left(W_B^{-2} \int_0^1 \ddot{\psi}_B(\beta^* + tb)dt \right) : B \cap S = \varnothing, b \in \mathscr{C}(\xi, S), \varphi_0(b) \le \eta^* \right\} \quad (45)$$

as a restricted upper eigenvalue, where $\lambda_{\max}(M)$ is the largest eigenvalue of matrix M, $B \subseteq \{1,\ldots, p\}$, and $\psi_B(\beta)$ and W_B are the restrictions of the Hessian of (1) and the weight operator $W = \mathrm{diag}(w_1,\ldots, w_p)$ to $\mathbb{R}^B$.

Theorem 19

Let β^* *be a target vector,* $S \supseteq \{j : \beta_j^* \neq 0\}$, $\hat{\beta}$ *be the weighted Lasso estimator (2), and* z_k^* *be the* ℓ_∞*-noise level as in (5). Let* $0 \le \eta^* \le 1$, $\phi_{1,S}(b) = |b_S|_1/|S|$, ϕ_0 *be a quasi star-shaped function, and* $F(\xi, S; \phi_0, \phi)$ *be the GIF in (14). Then, in the event (15),*

$$\#\{j : \hat{\beta}_j \neq 0, j \notin S\} < d_1 = \min\left\{ m \ge 1 : \frac{m}{\kappa_+(m)} > \frac{e^\eta \xi^2 |S|}{F(\xi, S; \varphi_0, \varphi_{1,S})} \right\}.$$

It follows from the Cauchy-Schwarz inequality that $\kappa_+(m)$ is sub-additive, $\kappa_+(m_1 + m_2) \le \kappa_+(m_1) + \kappa_+(m_2)$, so that $m/\kappa_+(m)$ is non-decreasing in m. For GLM, lower bounds for the GIF and probability upper bounds for z_k^* can be found in Subsection 3.2. For $S = \{j : \beta_j^* \neq 0\}$. Theorem 19 gives an upper bound for the false negative.

In linear regression, upper bounds for the false negative of the Lasso or concave penalized LSE can be found in Zhang and Huang (2008) and Zhang (2010a) under a sparse Riesz condition (SRC). We now extend their results to the Lasso for the more general convex minimization problem (1). For this purpose, we strengthen (18) to

$$e^{-\varphi_0(b)} \sum{}^* \le \ddot{\psi}(\beta^* + b) \le e^{\varphi_0(b)} \sum{}^*, \quad \forall b \in \mathscr{C}(\xi, S), \varphi_0(b) \le \eta^*, \quad (46)$$

and assume the following SRC: for certain constants $\{c_*, c^*\}$, integer d^*, $0 < \alpha < 1$, $0 < \eta \le \eta^* \le 1$, all $A \supset S$ with $|A| = d^*$, and all $u \in \mathbb{R}^A$ with $|u| = 1$,

$$c_* \le \langle u, \ddot{\psi}_A(\beta^*) u \rangle \le c^*, \frac{|S|}{2(1-\alpha)} \left(\frac{e^{2\eta} c^*}{c_*} + 1 - 2\alpha \right) \le d^*. \quad (47)$$

Theorem 20

Let $\hat{\beta}$ *be the Lasso estimator (2) with* $w_j = 1$ *for all* j, β^* *be a target vector,* $S \supseteq \{j:\beta_j^* \neq 0\}$, *and* z_k^* *be the* ℓ_∞*-noise level as in (5). Let* ϕ_0 *be a quasi star-shaped function, and* $F(\xi,S;\phi_0,\phi)$ *be the GIF in (14). Suppose (46) and (47) hold. Let* d_1 *be the integer satisfying* $d_1 - 1 \leq |S|(e^{2\eta}c^*/c_* - 1)/(2 - 2\alpha) < d_1$. *Then,*

$$\#\{j:\hat{\beta}_j \neq 0, j \notin S\} < d_1$$

when $z_0^* + \xi z_1^* \leq (\xi - 1)\lambda$, $\lambda + z_0^* \leq \eta e^{-\eta} F(\xi, S; \varphi_0, \varphi_0)$, *and*

$$\max_{A \supset S, |A| \leq d_1} |(\sum^*)_{AA}^{-1/2} \ell_A(\beta^*)|_2 \leq e^{-\eta}\alpha\lambda\sqrt{d_1/c^*}.$$

Theorems 19 and 20 use different sets of conditions to derive dimension bounds since different analytical approaches are used. These sets of conditions do not imply each other. In the most optimistic case, the SRC (47) allows $d^* = d_1 + |S|$ to be arbitrary close to $|S|$ when $e^{2\eta}c^*/c_* \approx 1$, while Theorem 19 requires $d_1 \geq |S|$ when $\kappa_+(m) \geq 1$ and $F(\xi,S;\phi_0,\phi_{1,S}) \leq 1$ (always true for Σ^* with 1 in the diagonal).

7. Discussion

In this paper, we studied the estimation, prediction, selection and sparsity properties of the weighted and adaptive ℓ_1-penalized estimators in a general convex loss formulation. We also studied concave regularization in the form of recursive application of adaptive ℓ_1-penalized estimators.

We applied our general results to several important statistical models, including linear regression and generalized linear models. For linear regression, we extend the existing results to weighted and adaptive Lasso. For the GLMs, the ℓ_q,$q \geq 1$ error bounds for a general $q \geq 1$ for the GLMs are not available in the literature, although ℓ_1 and ℓ_2 bounds have been obtained under different sets of conditions respectively in van de Geer (2008) and]citeNegahbanRWY10. Our fixed-sample analysis provides explicit definition of constant factors in an explicit neighborhood of a target. Our oracle inequalities yields even sharper results for multistage recursive application of adaptive Lasso based on a suitable concave penalty. The results on the sparsity of the solution to the ℓ_1-penalized convex minimization problem is based on a new approach.

An interesting aspect of the approach taken in this paper in dealing with general convex losses such as those for the GLM is that the conditions imposed on the Hessian naturally "converge" to those for the linear regression as the convex loss "converges" to a quadratic form.

A key quantity used in the derivation of the results is the generalized invertibility factor (14), which grow out of the idea of the ℓ_2 restricted eigenvalue but improves upon it. The use of GIF yields sharper bounds on the estimation and prediction errors. This was discussed in detail in the context of linear regression in Ye and Zhang (2010).

We assume that the convex function $\psi(\cdot)$ is twice differentiable. Although this assumption is satisfied in many important and widely used statistical models, it would be interesting to

extend the results obtained in this paper to models with less smooth loss functions, such as those in quantile regression and support vector machine.

Acknowledgments

The work of Jian Huang is supported in part by the National Institutes of Health (NIH Grants R01CA120988 and R01CA142774) and the National Science Foundation (NSF Grant DMS-08-05670). The work of Cun-Hui Zhang is supported in part by the National Science Foundation (NSF Grants DMS-0906420 and DMS-1106753) and the National Security Agency (NSA Grant H98230-11-1-0205).

References

Bickel PJ, Ritov Y, Tsybakov A. Simultaneous analysis of Lasso and Dantzig selector. Annals of Statistics. 2009; 37(4):1705–1732.

Bregman LM. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Computational Mathematics and Mathematical Physics. 1967; 7:200–217.

Bunea F, Tsybakov A, Wegkamp MH. Sparsity oracle inequalities for the Lasso. Electronic Journal of Statistics. 2007; 1:169–194.

Candes EJ, Tao T. Decoding by linear programming. IEEE Trans on Information Theory. 2005; 51:4203–4215.

Candes EJ, Tao T. The dantzig selector: statistical estimation when. p is much larger than n (with discussion). Annals of Statistics. 2007; 35:2313–2404.

Chen S, Donoho DL, Saunders MA. Atomic decomposition by basis pursuit. SIAM J Sci Comput. 1998; 20:33–61.

Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association. 2001; 96:1348–1360.

Greenshtein E, Ritov Y. Persistence in high–dimensional linear predictor selection and the virtue of overparametrization. Bernoulli. 2004; 10:971–988.

Huang J, Ma S, Zhang CH. Adaptive lasso for sparse high-dimensional regression models. Statistica Sinica. 2008; 18:1603–1618.

Hunter DR, Li R. Variable selection using mm algorithms. Annals of Statistics. 2005; 33:1617–1642. [PubMed: 19458786]

Koltchinskii V. The dantzig selector and sparsity oracle inequalities. Bernoulli. 2009; 15:799–828.

McCullagh, P.; Nelder, JA. Generalized Linear Models. Chapmann & Hall; 1989.

Meier L, Bühlmann P. Smoothing ℓ_1-penalized estimators for high-dimensional time-course data. Electronic Journal of Statistics. 2007; 1:597–615.

Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. Annals of Statistics. 2006; 34:1436–1462.

Meinshausen N, Yu B. Lasso-type recovery of sparse representations for high-dimensional data. Annals of Statistics. 2009; 37:246–270.

Negahban, S.; Ravikumar, P.; Wainwright, MJ.; Yu, B. Technical Report arXiv:1010.2731, arXiv. 2010. A unified framework for high-dimensional analysis of *m*-estimators with decomposable regularizer.

Nielsen F, Nock R. On the centroids of symmetrized bregman divergences. CoRR. 2007 abs/0711.3242.

Ravikumar P, Wainwright MJ, Raskutti G, Yu B. Model selection in gaussian graphical models: High-dimensional consistency of ℓ_1-regularized mle. Advances in Neural Information Processing Systems (NIPS). 2008; 21

Rothman AJ, Bickel PJ, Levina E, Zhu J. Sparse permutation invariant covariance estimation. Electronic Journal of Statistics. 2008; 2:494–515.

Städler N, Bühlmann P, van de Geer S. ℓ_1-penalization for mixture regression models (with discussion). Test. 2010; 19(2):209–285.

Sun, T.; Zhang, C-H. Technical Report arXiv:1104.4595, arXiv. 2011. Scaled sparse linear regression.

Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 1996; 58:267–288.

Tibshirani R, Taylor J. The solution path of the generalized lasso. The Annals of Statistics. 2011; 39:1335–1371.

Tropp JA. Just relax: convex programming methods for identifying sparse signals in noise. IEEE Transactions on Information Theory. 2006; 52:1030–1051.

van de Geer, S. Technical Report 140. ETH Zurich; Switzerland: 2007. The deterministic lasso.

van de Geer S. High–dimensional generalized linear models and the lasso. Annals of Statistics. 2008; 36:614–645.

van de Geer S, Bühlmann P. On the conditions used to prove oracle results for the lasso. Electronic Journal of Statistics. 2009; 3:1360–1392.

Wainwright MJ. Sharp thresholds for noisy and high–dimensional recovery of sparsity using ℓ_1–constrained quadratic programming (lasso). IEEE Transactions on Information Theory. 2009; 55:2183–2202.

Ye F, Zhang CH. Rate minimaxity of the lasso and dantzig selector for the ℓ_q loss in ℓ_r balls. Journal of Machine Learning Research. 2010; 11:3481–3502.

Zhang C-H. Least squares estimation and variable selection under minimax concave penalty. Mathematisches Forschungsintitut Oberwolfach: Sparse Recovery Problems in High Dimensions. 2009; 3

Zhang CH. Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics. 2010a; 38:894–942.

Zhang CH, Huang J. The sparsity and bias of the Lasso selection in high-dimensional linear regression. Annals of Statistics. 2008; 36(4):1567–1594.

Zhang T. Analysis of multi-stage convex relaxation for sparse regularization. Journal of Machine Learning Research. 2010b; 11:1087–1107.

Zhang T. Adaptive forward-backward greedy algorithm for learning sparse representations. IEEE Transactions on Information Theory. 2011a; 57:4689–4708.

Zhang, T. Technical Report arXiv:1106.0565, arXiv. 2011b. Multi-stage convex relaxation for feature selection.

Zhao P, Yu B. On model selection consistency of Lasso. Journal of Machine Learning Research. 2006; 7:2541–2567.

Zou H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association. 2006; 101:1418–1429.

Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models. Annals of Statistics. 2008; 36(4):1509–1533. [PubMed: 19823597]

Appendix A

Proof of Lemma 1

Since $\dot{\psi}(\hat{\beta}) - \dot{\psi}(\beta^*) = z - \dot{\psi}(\beta^*) - g$, (3) implies

$$\Delta(\hat{\beta}, \beta^*) = \langle \hat{\beta}, z - \dot{\psi}(\beta^*) \rangle - \lambda |\widehat{W}\hat{\beta}|_1 - \langle \beta^*, z - \dot{\psi}(\beta^*) - g \rangle$$

and $|g_j| \le \hat{w}_j\lambda$. Thus, (7) follows from $|(z - \dot{\psi}(\beta^*)_j| \le \hat{w}_j\lambda$ and $\hat{w}_j \le w_j$ in S in Ω_0.

For (8), we have $h_{S^c} = \hat{\beta}_{S^c}$ and $\beta^*_{S^c} = 0$, so that in Ω_0 (3) gives

$$\begin{aligned}\Delta(\widehat{\beta},\beta^*)&=\langle\widehat{\beta}_{S^c},\{z-\dot{\psi}(\beta^*)\}_{S^c}\rangle-\lambda|\widehat{W}_{S^c}\widehat{\beta}_{S^c}|_1-\langle h_S,\{z-\dot{\psi}(\beta^*)-g\}_S\rangle\\&\le|W_{S^c}\widehat{\beta}_{S^c}|_1(z_1^*-\lambda)+\langle h_S,g_S-\{z-\dot{\psi}(\beta^*)\}_S\rangle\\&\le|W_{S^c}\widehat{\beta}_{S^c}|_1(z_1^*-\lambda)+|h_S|_1(z_0^*+|w_S|_\infty\lambda).\end{aligned}$$

This gives (8). Since $\Delta(\widehat{\beta},\beta^*)>0$, $h\in\mathscr{C}(\xi,S)$ when $(|w_S|_\infty\lambda+z_0^*)/(\lambda-z_1^*)\le\xi$. For $j\notin S$, $h_j(\psi(\beta+h)-\psi(\beta))_j=\widehat{\beta_j}(z-\psi(\beta^*)-g)_j\le|\widehat{\beta_j}|(w_j\lambda-g_j)\le 0$.

Proof of Theorem 4

Let $h=\widehat{\beta}-\beta^*$. Since $\psi(\beta)$ is a convex function,

$$t^{-1}\Delta(\beta^*+th,\beta^*)=\frac{\partial}{\partial t}\left\{\psi(\beta^*+th)-t\langle h,\dot{\psi}(\beta^*)\rangle\right\}$$

is an increasing function of t. For $0\le t\le 1$ and in the event Ω_1, (8) implies

$$t^{-1}\Delta(\beta^*+th,\beta^*)\le\Delta(h+\beta^*,\beta^*)<(|w_S|_\infty\lambda+z_0^*)|h_S|_1.$$

By (9) and (14), $F(\xi,S;\phi_0,\phi_0)\le\Delta(\beta^*+h,\beta^*)e^{\phi_0(th)}/\{t|h_S|_1\phi_0(th)\}$ for $\phi_0(th)\le\eta^*$. Thus, for $\phi_0(th)\le\min\{\eta^*,\phi_0(h)\}$ and in the event Ω_1,

$$\varphi_0(th)e^{-\varphi_0(th)}\le\frac{\Delta(\beta^*+th,\beta^*)}{t|h_S|_1F(\xi,S;\varphi_0,\varphi_0)}\le\frac{|w_S|_\infty\lambda+z_0^*}{F(\xi,S;\varphi_0,\varphi_0)}\le\eta e^{-\eta}.$$

If $\eta^*<\phi_0(h)$, the above inequality at $\phi_0(th)=\eta^*$ would give $\eta^*e^{-\eta^*}<\eta e^{-\eta}$, which contradicts to $\eta\le\eta^*\le 1$. Thus, $\eta^*\ge\phi_0(h)$ and $\phi_0(th)e^{-\phi_0(th)}\le\eta e^{-\eta}$ for all $0\le t\le 1$. This implies $\phi_0(h)\le\eta\le\eta^*$. Another application of (8) yields

$$\varphi(h)\le\frac{\Delta(\beta^*+h,\beta^*)e^{\varphi_0(h)}}{F(\xi,S;\varphi_0,\varphi)|h_S|_1}\le\frac{(|w_S|_\infty\lambda+z_0^*)e^{\eta}}{F(\xi,S;\varphi_0,\varphi)}.$$

We obtain (17) by applying (16) with $\phi=\phi_{1,S}$ to the right-hand side of (8).

Proof of Lemma 8

i. Since $\dot{\psi}(\beta)=\sum_{i=1}^n x^i\dot{\psi}_0(x^i\beta)/n$ by (23),

$$\begin{aligned}\mathrm{E}_\beta\exp\left\{\tfrac{n}{\sigma^2}\langle b,z-\dot{\psi}(\beta)\rangle\right\}&=\exp\left[\sum_{i=1}^n\frac{\psi_0(x^i(\beta+b))-\psi_0(x^i\beta)-(x^ib)\dot{\psi}_0(x^i\beta)}{\sigma^2}\right]\\&=\exp\left[\sum_{i=1}^n\int_0^1\frac{(x^ib)^2\ddot{\psi}_0(x^i(\beta+tb))}{\sigma^2}(1-t)dt\right].\end{aligned}\quad(48)$$

This and (24) imply that for $M_1|Xb|_\infty\le\eta_0$,

$$\mathrm{E}_{\beta^*}\exp\left\{\frac{n}{\sigma^2}\langle b, z-\dot{\psi}(\beta^*)\rangle\right\} \le \exp\left[\frac{ne^{\eta_0}\langle b, \sum^* b\rangle}{2\sigma^2}\right]. \quad (49)$$

Since $\max_{k=0,1} z_k^*/\lambda_k = \max_j t_j^{-1}|z_j - \dot{\psi}_j(\beta^*)$ by (5),

$$\begin{aligned}\mathrm{P}_{\beta^*}\left\{\max_{k=0,1} z_k^*/\lambda_k > 1\right\} &\le \sum_{j=1}^{p} \mathrm{P}_{\beta^*}\left\{|z_j - \dot{\psi}_j(\beta^*)| > t_j\right\} \\ &\le \sum_{j=1}^{p} \mathrm{E}_{\beta^*}\exp\left\{\tfrac{n}{\sigma^2} b_j |z_j - \dot{\psi}_j(\beta^*)| - \tfrac{n}{\sigma^2} b_j t_j\right\}\end{aligned}$$

with $b_j = e^{-\eta_0} t_j / \sum_{jj}^*$. Since $M_1 \max_{ij}|x_{ij}| b_j \le \eta_0$, (49) gives

$$\mathrm{P}_{\beta^*}\left\{\max_{k=0,1} z_k^*/\lambda_k > 1\right\} \le \sum_{j=1}^{p} 2\exp\left(-\frac{ne^{-\eta_0} t_j^2}{2\sigma^2 \sum_{jj}^*}\right).$$

ii. If (30) holds, we simply replace $\ddot{}_0(x^i(\beta + tb))$ by c_0 in (48). The rest is simpler and omitted.

Proof of Theorem 9

(i) Since $F^*(\xi, S; \phi)$ in (25) is a lower bound of $F(\xi, S; \phi_0, \phi)$ in (14), (32) follows from Theorem 4 with $\phi_0(b) = M_2|b|_2$. The probability statement follows from Lemma 8. (ii) Since (18) holds for the $\phi_0(b)$ in (26), we are allowed to use $F_*(\xi, S) = F_0(\xi, S; \phi_0)$ in Corollary 7. The condition $\eta^* = \infty$ is used since $\phi_0(b)$ does not control $M_1|Xb|_\infty$. (iii) We are also allowed to use $\phi_0(b) = M_3|b_S|_1$ in (28) due to $M_1|Xb|_\infty \le \phi_0(b)$.

Proof of Theorem 14

Let $h = \hat{\beta} - \beta^*$, $w_j = \hat{w}_j$ and $S = \{j : |\hat{\beta_j}| > \gamma_0\lambda\} \cup S_0$. For $j \notin S$, $w_j = \dot{\rho_\lambda}(|\tilde{\beta_j}|)/\lambda \ge \{\dot{\rho_\lambda}(0+) - \kappa\gamma_0\lambda\}/\lambda = 1 - \kappa\gamma_0$, so that $z_1^* = |W_{S^c}^{-1}\{z - \dot{\psi}(\beta^*)\}_{S^c}|_\infty \le \lambda_0/(1-\kappa\gamma_0) = \lambda/A$. We also have $z_0^* \le |z - \dot{\psi}(\beta^*)|_\infty \le \lambda_0 = (1-\kappa\gamma_0)/\lambda/A$. Since $|\hat{w}|_\infty \le 1$, these bounds for z_0^* and z_1^* yield

$$\frac{|\hat{w}_S|_\infty\lambda + z_0^*}{\lambda - z_1^*} \le \frac{\lambda + (1-\kappa\gamma_0)\lambda/A}{\lambda - \lambda/A} = \frac{A + 1 - \kappa\gamma_0}{A - 1} \le \xi.$$

Thus, since $|g_j| \le \hat{w}_j\lambda$ in (8), Lemma 1 provides

$$h \in \mathscr{C}(\xi, S), \Delta(\beta^* + h, \beta^*) \le |h_S|_2(|\hat{w}_S|_2\lambda + |\{z - \dot{\psi}(\beta^*)\}_S|_2)$$

Since $|S \backslash S_0| \le |(\tilde{\beta} - \beta^*)_{S_0^c}|_2^2/(\gamma_0^2\lambda^2) \le \ell^*$, we have by (36)

$$|w_S|_\infty\lambda + z_0^* \le \lambda + \lambda_0 = \lambda_0\{1 + A/(1-\kappa\gamma_0)\} \le F(\xi, S; \varphi_0, \varphi_0)\eta e^{-\eta}.$$

Thus, $\phi_0(h) \le \eta$ by (16), so that by (34) and (36),

$$e^{-\eta}F_*|h_S|_2|h|_2 \le \Delta(\beta^*+h,\beta^*) \le |h_S|_2(|\widehat{w}_S|_2\lambda+|\{z-\dot{\psi}(\beta^*)\}_S|_2).$$

Since $|h_S| = 0$ implies $h = 0$ for $h \in \mathscr{C}(\xi,S)$, we find

$$e^{-\eta}F_*|h|_2 \le |\widehat{w}_S|_2\lambda+|\{z-\dot{\psi}(\beta^*)\}_S|_2. \quad (50)$$

Since $\widehat{w}_j\lambda=\dot{\rho}_\lambda(|\tilde{\beta}_j|) \le \dot{\rho}_\lambda(|\beta_j^*|)+\kappa|\tilde{\beta}_j-\beta_j^*|$, we have

$$|\widehat{w}_S|_2\lambda \le |\dot{\rho}_\lambda(|\beta^*_{S_0}|)|_2+\kappa|\tilde{\beta}-\beta^*|_2.$$

Since $|z - \psi(\dot{\beta}^*)|_\infty \le \lambda_0 = (1 - \kappa\gamma_0)\lambda/A$ and $|\tilde{\beta}_j-\beta_j^*|=|\tilde{\beta}_j| \ge \gamma_0\lambda$ for $j \in S\backslash S_0$,

$$\begin{aligned}|\{z-\dot{\psi}(\beta^*)\}_S|_2 &\le |\{z-\dot{\psi}(\beta^*)\}_{S_0}|_2+|S\backslash S_0|^{1/2}(1-\kappa\gamma_0)\lambda/A\\ &\le |\{z-\dot{\psi}(\beta^*)\}_{S_0}|_2+|\tilde{\beta}-\beta^*|_2(1-\kappa\gamma_0)/(\gamma_0 A).\end{aligned}$$

Inserting the above inequalities into (50), we find that

$$e^{-\eta}F_*|\widehat{\beta}-\beta^*|_2 \le |\dot{\rho}_\lambda(|\beta^*_{S_0}|)|_2+|\{z-\dot{\psi}(\beta^*)\}_{S_0}|_2+\left(\kappa+\frac{1}{\gamma_0 A}-\frac{\kappa}{A}\right)|\tilde{\beta}-\beta^*|_2.$$

The probability statement follows directly from (35) with the union bound.

Proof of Theorem 15

Let $R^{(\ell)}$ be as in Remark 16. For $|z - \psi(\dot{\beta}^*)|_\infty \le \lambda_0$, (16) of Theorem 4 gives $|\hat{\beta}^{(0)} - \beta^*|_2 \le e^\eta(\lambda + \lambda_0)/F(\xi_0,S_0;\phi_0, |\cdot|_2) = R^{(0)}$. Under conditions (38) and (40), we have $R^{(\ell)} \le \gamma_0\lambda\sqrt{\ell^*}$ for all $\ell \ge 0$. We prove (39) by induction. We have already proved (39) for $\ell = 0$. For $\ell \ge 1$, we let $\tilde{\beta} = \hat{\beta}^{(\ell-1)}$ and apply Theorem 14: $|\hat{\beta}^{(\ell)} - \beta^*|_2 \le (1 - r_0)R^{(\infty)} + r_0R^{(\ell-1)} = R^{(\ell)}$. The probability statement follows directly from (35) with the union bound.

Proof of Theorem 18

Let $z \tilde{=} z - \psi(\dot{\beta}^*)$ and λ be fixed. Consider

$$\widehat{\beta}(\lambda,t)=\arg\min_\beta\left\{\psi(\beta)-\langle\beta,\dot{\psi}(\beta^*)+t\tilde{z}\rangle+t\lambda\sum_{j=1}^p\widehat{w}_j|\beta_j|:\beta_{S^c}=0\right\} \quad (51)$$

as an artificial path for $0 \le t \le 1$. For each t, the KKT conditions for $\hat{\beta}(\lambda, t)$ are

$$g_S(\lambda,t)=t\lambda\widehat{W}_S u_S(\lambda,t),\ u_j(\lambda,t)\begin{cases}=\mathrm{sgn}(\widehat{\beta}_j(\lambda,t)) & \forall\widehat{\beta}_j(\lambda,t)\neq 0\\ \in[-1,1], & \forall j\in S,\end{cases}$$

where $g(\lambda, t) = -\dot{\psi}(\hat{\beta}(\lambda, t)) + \dot{\psi}(\beta^*) + t\tilde{z}$. Since (51) is constrained to $\beta_{S^c} = 0$ and both the error $\tilde{z}$ and the penalty level λ are scaled with t, Theorem 4 with $\xi = 0$ yields

$$\varphi_0(\widehat{\beta}(\lambda,t)-\beta^*)\le\eta_t\to 0\,\text{with}\,\eta_t e^{-\eta_t}=t\eta e^{-\eta},\forall 0<t\le 1. \quad (52)$$

Let $S_t = \{j : \hat{\beta}_j(\lambda, t) \neq 0\}$. Applying the differentiation operator $D = (\partial/\partial t)$ to the KKT conditions, we find that almost everywhere in t,

$$(Dg)_{S_t}(\lambda,t)=\tilde{z}_{S_t}-\ddot{\psi}_{S_t}(\widehat{\beta}(\lambda,t))\{(D\widehat{\beta})(\lambda,t)\}_{S_t}=\lambda\widehat{W}_{S_t}u_{S_t}(\lambda,t).$$

It follows that

$$(D\widehat{\beta})_{S_t}(\lambda,t)=\{\ddot{\psi}_{S_t}(\widehat{\beta}(\lambda,t))\}^{-1}\{\tilde{z}_{S_t}-\lambda\widehat{W}_{S_t}u_{S_t}(\lambda,t)\} \quad (53)$$

and with an application of the chain rule,

$$(Dg)_{S^c}(\lambda,t)=\tilde{z}_{S^c}-\ddot{\psi}_{S^c,S_t}(\widehat{\beta}(\lambda,t))\{\ddot{\psi}_{S_t}(\widehat{\beta}(\lambda,t))\}^{-1}\{\tilde{z}_{S_t}-\lambda\widehat{W}_{S_t}u_{S_t}(\lambda,t)\}.$$

Since $g(\lambda, t)$ is almost differentiabe and $\hat{\beta}(\lambda, 0+) = \beta^*$, we have $g(\lambda, 0+) = 0$ and

$$g_{S^c}(\lambda,1-)=\int_0^1\left[\tilde{z}_{S^c}-\ddot{\psi}_{S^c,S_t}(\widehat{\beta}(\lambda,t))\{\ddot{\psi}_{S_t}(\widehat{\beta}(\lambda,t))\}^{-1}\{\tilde{z}_{S_t}-\lambda\widehat{W}_{S_t}u_{S_t}(\lambda,t)\}\right]dt.$$

Thus, (52), (41), and (42) imply

$$|\widehat{W}_{S^c}^{-1}g_{S^c}(\lambda,1-)|_\infty\le|\widehat{W}_{S^c}^{-1}\tilde{z}_{S^c}|_\infty+\kappa_1|\tilde{z}_S|_\infty+\kappa_0\lambda|u_{S_t}(\lambda,t)|_\infty,$$

which is smaller than λ in the event in (43). Thus, since $\ddot{\psi}_S(\hat{\beta}(\lambda, 1-))$ is of full rank, $\hat{\beta}(\lambda, 1-)$ is the unique solution of the KKT conditions (3) for $\hat{\beta}$. This completes the proof of part (i).

For part (ii), we observe that (44) implies $S=\{j:\beta_j^*\neq 0\}$. Since $\hat{\beta}(\lambda, 0+) = \beta^*$, there exists $t_1 > 0$ such that $u_S(\lambda,t)=\mathrm{sgn}(\beta_S^*)$ for all $0 < t < t_1$. By (52), $\hat{\beta}(\lambda, t) \in \mathscr{B}_0$ for $0 < t < t_1$. It follows from (53) and (44) that

$$|(D\widehat{\beta})_S(\lambda,t)|_\infty\le\|\{\ddot{\psi}_{S_t}(\widehat{\beta}(\lambda,t))\}^{-1}\|_\infty|\tilde{z}_S-\lambda\widehat{W}_S\mathrm{sgn}(\beta_S^*)|_\infty<\min_{j\in S}|\beta_j^*|-\varepsilon_1$$

for $0 < t < t_1$ and some $\varepsilon_1 > 0$. Since $\hat{\beta}(\hat{\lambda}, 0+) = \beta^*$, this implies $|\hat{\beta}_S(\lambda, t) - \beta^*_S|_\infty < \min_{j \in S} |\beta^*_j| - \varepsilon_1$ for all $0 < t < t_1 \wedge 1$. It follows that $\operatorname{sgn}(\hat{\beta}(\hat{\lambda}, t)) = \operatorname{sgn}(\beta^*)$ for $0 < t \le 1$ by the continuity of $\hat{\beta}(\hat{\lambda}, t)$ in t, that is, $t_1 = 1$. Consequently, conditions (41), and (42) are only needed for the smaller class $\mathscr{A}_0$ in the proof of part (i). This gives $\hat{\beta}(\hat{\lambda}, 1) = \hat{\beta}$ and completes the proof of part (ii).

Finally, in part (iii), $F_0(\xi, S; \phi_0, \phi_0)$ is simply replaced by its lower bounds with the respective ϕ_0.

Proof of Theorem 19

Suppose the event Ω_1 in (15) happens, so that $\hat{w}_j \ge w_j$ for $j \notin S$ and the conclusion of Theorem 4 hold. Let $h = \hat{\beta} - \beta^*$ and $\widehat{\sum} = \int_0^1 \ddot{\psi}(\beta^* + xh)dx$. It follows from (1) that $\Sigma\hat{h} = \dot{\psi}(\beta^* + h) - \dot{\psi}(\beta^*) = \dot{\ell}(\hat{\beta}) - \dot{\ell}(\beta^*)$. By the KKT conditions (3),

$$|(\widehat{\sum} h)_j| = |(\dot{\ell}(\hat{\beta}) - \dot{\ell}(\beta^*))_j| \ge \hat{w}_j \lambda - z_j \ge w_j(\lambda - z_1^*) > 0, \quad j \notin S.$$

Let $B \subseteq \{j \notin S : \hat{\beta}_j \neq 0\}$ with $|B| \le d_1$. It follows from Theorem 4 that $\phi_0(h) \le \eta \le \eta^*$, so that (45) implies $\max_{|u|_2 = 1} |(W^{-1} \widehat{\sum}^{1/2} u)_B|_2^2 = \lambda_{\max}(W_B^{-2} \widehat{\sum}_B) \le \kappa_+(d_1)$. Thus, by the definition of $\Delta(\beta, \beta^*)$ in (4),

$$(\lambda - z_1^*)^2 |B| \le |(W^{-1} \widehat{\sum} h)_B|_2^2 \le \kappa_+(d_1) \langle h, \widehat{\sum} h \rangle = \kappa_+(d_1) \Delta(\beta^* + h, \beta^*).$$

This and the prediction bound in Theorem 4 yield

$$|B| \le \frac{\kappa_+(d_1)\Delta(\beta^* + h, \beta^*)}{(\lambda - z_1^*)^2} \le \frac{\kappa_+(d_1) e^\eta (|w_S|_\infty \lambda + z_0^*)^2 |S|}{(\lambda - z_1^*)^2 F(\xi, S; \varphi_0, \varphi_{1,S})} \le \frac{\kappa_+(d_1) e^\eta \xi^2 |S|}{F(\xi, S; \varphi_0, \varphi_{1,S})} < d_1.$$

Since all subsets $B \subseteq \{j \notin S : \hat{\beta}_j \neq 0\}$ with $|B| \le d_1$ satisfies $|B| < d_1$, it must hold that $\#\{j \notin S : \hat{\beta}_j \neq 0\} < d_1$.

Proof of Theorem 20

Let $z \tilde{=} z - \dot{\psi}(\beta^*) = -\dot{\ell}(\beta^*)$ and $\hat{\beta}(\hat{\lambda}, t)$ be the artificial estimator in (51) with $\hat{w}_j = 1$, and $h(\lambda, t) = \hat{\beta}(\hat{\lambda}, t) - \beta^*$. Let $\lambda_* \le \lambda^*$ be penalty levels satisfying

$$[\lambda_*, \lambda^*] \subseteq \cap_{0 < t \le 1} \left\{ \lambda : \varphi_0(h(\lambda, t)) \le \eta, h(\lambda, t) \in \mathscr{C}(\xi, S), |(\sum^*)^{-1/2} \tilde{z}|_2 \le \frac{\alpha \lambda \sqrt{d_1}}{e^\eta \sqrt{c^*}} \right\}. \quad (54)$$

We pick such an interval $[\lambda_*, \lambda^*]$ containing the penalty level λ of concern in the theorem. This is allowed by Lemma 1 and Theorem 4. We first prove the stronger conclusion

$$\max_{\lambda_* \le \lambda \le \lambda^*} \max_{0<t\le 1} \#\{j:\hat{\beta}_j(\lambda,t) \neq 0, j \notin S\} < d_1 \quad (55)$$

under the additional assumption

$$\min_{\lambda_* \le \lambda \le \lambda^*} \min_{0<t\le 1} \#\{j:\hat{\beta}_j(\lambda,t) \neq 0, j \notin S\} \le d_1. \quad (56)$$

Let $g(\lambda, t) = tz \tilde{+} \dot{\psi}(\beta^*) - \dot{\psi}(\hat{\beta}(\lambda, t))$ be the negative gradient at $\hat{\beta}(\lambda, t)$ in (51). By the KKT conditions for (51), $\mathrm{sgn}(\hat{\beta}_j(\lambda, t\pm)) \neq 0$ implies $|g(\lambda, t)| = t\lambda$. Thus, (56) implies the existence of $\lambda \in [\lambda_*, \lambda^*]$, $t_1 \in (0, 1]$, and $A_1 \subset \{1, \ldots, p\}$ satisfying

$$\{j:\mathrm{sgn}(\hat{\beta}_j(\lambda,t_1)) \neq 0\} \cup S \subseteq A_1 \subseteq \{j:|g(\lambda,t_1)|=t_1\lambda\} \cup S,\ |A_1| \le d_1+|S|. \quad (57)$$

Moreover, if $\max_{\lambda_* \le \lambda \le \lambda^*} \max_{0<t\le 1} \#\{j : \hat{\beta}_j(\lambda, t) \neq 0, j \notin S\} \ge d_1$, then by the continuity of $\hat{\beta}(\lambda, t)$, it would be possible to restrict (57) to $|A_1| = d_1 + |S|$ with some different $\lambda \in [\lambda_*, \lambda^*]$ and $t_1 \in (0, 1]$. Therefore, it suffices to deny this possibility by proving $|A_1| < d_1 + |S|$ based on (57) and (54). Let $A_0 = A_1 \backslash S$. We prove $|A_0| < d_1$, which is equivalent to $|A_1| < d_1 + |S|$.

Let $v_{(A)} = (v_j I\{j \in A\}, j \in A_1)' \in \mathbb{R}^{A_1}$ and $v_A = (v_j, j \in A)' \in \mathbb{R}^A$ for all vectors $v = (v_1, \ldots, v_p)'$. Let $h = h(\lambda, t_1)$, $\widehat{\sum} = \int_0^1 \ddot{\psi}(\beta^* + xh)dx$, and $g = g(\lambda, t_1) = t_1 z \tilde{+} \dot{\psi}(\beta^*) - \dot{\psi}(\beta^* + h) = t_1 z \tilde{-} \Sigma\hat{h}$. Since $h_{A_1^c} = 0$, $\widehat{\sum}_{A_1}^{-1} g_{(A_1)} = t_1 \widehat{\sum}_{A_1}^{-1} \tilde{z}_{A_1} - \widehat{\sum}_{A_1}^{-1} (\widehat{\sum} h)_{A_1} = t_1 \widehat{\sum}_{A_1}^{-1} \tilde{z}_{A_1} - h_{A_1}$. Thus, since $g_j = t_1\lambda\,\mathrm{sgn}(h_j)$ for $j \in A_0$ by the KKT conditions,

$$\langle g_{(A_0)}, \widehat{\sum}_{A_1}^{-1} g(A_1)\rangle = t_1 \langle g_{(A_0)}, \widehat{\sum}_{A_1}^{-1} \tilde{z}_{A_1}\rangle - \langle g_{(A_0)}, h_{A_1}\rangle \le t_1 \langle g_{(A_0)}, \widehat{\sum}_{A_1}^{-1} \tilde{z}_{A_1}\rangle.$$

Since $|\widehat{\sum}_{A_1}^{-1/2} g_{(A_0)}|_2^2 + |\widehat{\sum}_{A_1}^{-1/2} g_{(A_1)}|_2^2 = |\widehat{\sum}_{A_1}^{-1/2} g_{(S)}|_2^2 + 2\langle g_{(A_0)}, \widehat{\sum}_{A_1}^{-1} g_{(A_1)}\rangle$, we have

$$|\widehat{\sum}_{A_1}^{-1/2} g_{(A_0)}|_2^2 + |\widehat{\sum}_{A_1}^{-1/2} g_{(A_1)}|_2^2 \le |\widehat{\sum}_{A_1}^{-1/2} g(s)|_2^2 + 2t_1 |\widehat{\sum}_{A_1}^{-1/2} g_{(A_0)}|_2 |\widehat{\sum}_{A_1}^{-1/2} \tilde{z}_{A_1}|_2.$$

By (54) and (46), $|\widehat{\sum}_{A_1}^{-1/2} \tilde{z}_{A_1}|_2 \le e^{\eta/2} |(\sum_{A_1}^*)^{-1/2} \tilde{z}_{A_1}|_2 \le \alpha\lambda \sqrt{|A_0|/(c^* e^\eta)}$, so that

$$(1-\alpha)|\widehat{\sum}_{A_1}^{-1/2} g_{(A_0)}|_2^2 + |\widehat{\sum}_{A_1}^{-1/2} g_{(A_1)}|_2^2 \le |\widehat{\sum}_{A_1}^{-1/2} g(s)|_2^2 + \alpha t_1^2 \lambda^2 |A_0|/(c^* e^\eta).$$

Moreover, since $|A_1| \le d_1 + |S| \le d^*$, it follows from (54), (46), and (47) that the eigenvalues of $\Sigma_{\hat{A}_1}$ all lie in the interval $c_* e^{-\eta}$ and $c^* e^\eta$. Thus, since $g_{A_0} = t_1\lambda\,\mathrm{sgn}(\hat{\beta}_{A_0})$,

$$\frac{(1-\alpha)t_1^2\lambda^2|A_0|}{c^*e^{\eta}}+\frac{t_1^2\lambda^2|A_0|+|g_S|_2^2}{c^*e^{\eta}}\le\frac{|g_S|_2^2}{c_*e^{-\eta}}+\frac{\alpha t_1^2\lambda^2|A_0|}{c^*e^{\eta}}.$$

Since $|g|_\infty \le t_1\lambda$, the above inequality gives by algebra the dimension bound

$$|A_0|\le\left(\frac{e^{2\eta}c^*/c_*-1}{2-2\alpha}\right)\frac{|g_S|_2^2}{t_1^2\lambda^2}\le\left(\frac{e^{2\eta}c^*/c_*-1}{2-2\alpha}\right)|S|<d_1.$$

This proves (55) under the additional assumption (56).

Now we prove (56). In the special case of $\phi_0(b) = 0$, the condition on λ in (54) is monotone so that we are allowed to pick $\lambda^* = \infty$. Since $\beta(\lambda, 1) = 0$ for very large λ, (56) holds automatically for $\phi_0(b) = 0$. By (46), this special case is equivalent to linear regression since the Hessian does not depend on β. The difference of the general model (1) from linear regression is that the condition $\lambda+z_0^*\le\eta e^{-\eta}F(\xi,S;\varphi_0,\varphi_0)$, which excludes large λ, is needed to prove $\phi_0(h(\lambda, t)) \le \eta$ by Theorem 4. To overcome this difficulty, we consider very small $t > 0$. Let $b = (\beta - \beta^*)/t$. By (51),

$$\begin{aligned}t^{-1}\{\widehat{\beta}(\lambda,t)-\beta^*\}&=\arg\min_b\left\{\psi(\beta^*+tb)-\langle tb,\dot{\psi}\,(\beta^*)+t\tilde{z}\rangle+t\lambda|\beta^*+tb|_1\right\}\\&=\arg\min_b\left\{\int_0^1(1-x)\,\langle tb,\ddot{\psi}(\beta^*+xtb)tb\rangle dx-t^2\langle b,\tilde{z}\rangle+t\lambda|\beta^*+tb|_1\right\}\\&=\arg\min_b\left\{\int_0^1(1-x)\,\langle b,\ddot{\psi}(\beta^*+xtb)b\rangle dx-\langle b,\tilde{z}\rangle+\lambda|\beta^*/t+b|_1\right\}.\end{aligned}$$

Let $S_0=\{j:\beta_j^*\neq 0\}$. Since $\lambda|\beta^*/t+b|_1-\lambda|\beta^*|_1/t\to\lambda\langle\mathrm{sgn}(\beta^*),b\rangle+\lambda|b_{S_0^c}|_1$ as $t \to 0+$, $t^{-1}\{\hat{\beta}(\lambda, t) - \beta^*\}$ converges (along a subsequence if necessary) to

$$\widehat{b}(\lambda)=\arg\min_b\left\{2^{-1}\langle b,\ddot{\psi}(\beta^*)b\rangle-\langle b,\tilde{z}\rangle+\lambda\langle\mathrm{sgn}(\beta^*),b\rangle+\lambda|b_{S_0^c}|_1\right\}.$$

Moreover, since $z \tilde{} \ddot{}(\beta^*)b(\hat{\lambda})$ is the negative gradient at $b(\hat{\lambda})$, we have

$$\{j:|g_j(\lambda,t)|=t\lambda,j\notin S\}\to\{j\notin S:(\tilde{z}-\ddot{\psi}(\beta^*)\widehat{b}(\lambda))_j=\lambda\mathrm{sgn}(\widehat{b}_j(\lambda))\}.\quad(58)$$

Since this limit does not depend on $\phi_0(\cdot)$, the dimension bound (55) in the special case of linear regression implies that the right-hand side of (58) contains a smaller number of elements than d_1. This gives (56) in the general case by (58) and completes the proof.